Exploring Students' Calibration in High School Mathematics: A Classroom Intervention

by

Behnaz Herbst

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Secondary Education
University of Alberta

Abstract

Through this study, I sought to explore the change in students' calibration in a Grade 11 mathematics classroom. The intention was to inquire about the extent to which students' self-estimations of task performance aligned with their actual task performance on classroom tests. Such alignment is generally referred to as metacognitive calibration, or calibration for brief. Calibration indicates how closely students' self-estimations of their task performance outcomes match their actual task performance outcomes in any field of study. Improving students' task performance outcomes (i.e., their grades)—although often a by-product of improved calibration—was not the intention of this study. To explain, in my attempt to study students' calibration and to gain insights into the potential change and development in students' calibration, I designed and used a pedagogical intervention in a Grade 11 mathematics classroom. However, influencing students' task performance outcomes was not a pursued goal of my study.

To conduct this research, I used a mixed methods methodology in three phases of pre-intervention, intervention, and post-intervention. The pre- and post-intervention phases involved the collection of only qualitative data, whereas the intervention phase involved the collection of both qualitative and quantitative data. My intervention consisted of four components: monitoring strategy training, curriculum-based classroom tests, students' calibration graphs, and students' reflection notecards. As such, I incorporated the elements of these four components into a classroom intervention that I merged with students' daily learning activities. Except for the final exam and the follow-up interviews, all my intervention activities were completed between February 3, 2016 and May 13, 2016 (see Appendix J). However, with ethics permission, the students made one last performance estimation on their final exam and the follow-up interviews were conducted after the end of the school year, in the first two weeks of July 2016.

My qualitative data sources consisted of student questionnaires, reflection notecards, and follow-up interviews, while the quantitative data sources consisted of students' task performance outcomes, task performance estimations, and students' individually prepared calibration graphs. I performed one round of statistical data analysis followed by three rounds of blended data analyses in which I combined qualitative and quantitative data to facilitate an in-depth analysis of the collected research data.

The findings revealed an optimistic view in which calibration was perceived to be malleable, improving incrementally in the context of the employed classroom intervention, while also challenging the prevalent assumption that calibration can only be understood and explained numerically. The findings have potential implications for mathematics education, particularly for high school mathematics teachers. The results demonstrated that although cognitive skills are clearly necessary for learning mathematics, being metacognitively calibrated provides mathematics students with an understanding and awareness of their subject-related strengths and weaknesses so they can allocate their time, attention, and study efforts accordingly. The findings also underscore the key role of students' self-perceptions in guiding their calibration and provide insights for future research on calibration.

Keywords: metacognition, calibration, confidence judgement, mathematics education

Preface

This dissertation is an original study by Behnaz Herbst. The research project, of which this dissertation is a part of, under the title of "A Classroom Study of the Development of Mathematics Students' Metacognitive Calibration", received ethics approval from the University of Alberta Research Ethics Board (Study ID Pro00057546) on September 28, 2015, and from the District School Board External Research Review Committee on January 12, 2016.

Acknowledgement

I would like to acknowledge the assistance and guidance which I received from my supervisor, Professor Florence Glanfield and my co-supervisor, Professor Gregory Thomas, without whose help I would not have been able to complete this dissertation. I would also like to thank my examining committee Professors David Anderson, Christina Rinaldi, Bill Dunn, and Janelle McFeetors for their support and constructive comments.

With utmost gratitude, I further acknowledge my family's unbounded love, dedication, and the sacrifices they made so that I could fully commit myself to my intense learning. Thank you, Andrew, Azadeh, Michael, and Christian! In loving memory of my dear father, I am also grateful to my beloved mother, a retired teacher, who kept me grounded by humorously reminding me that I was going to be the 4th sibling with a PhD!

Finally, I owe a debt of gratitude to the mathematics students I have taught since the 1990s, when I left engineering and became a mathematics teacher, and in particular, to 24 wonderful Grade 11 students who participated in my research. Thank you all for accompanying me on my lifelong journey of learning. I have learned much from you and have felt proud and privileged to serve as your teacher. This thesis is inspired by and written for you. Thank you!

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

Overview

- *Look Behnaz, how many sparrows do you see on that old sycamore?*
- *Will you bring my wallet? How much money do you think I have?*
- *How many rose buds are there on the rose bush by the pond?*
- *Do the opposite sides of this carpet look exactly the same to you?*
- *How much taller than the twins do you think you are now?*

For my math enthusiast father, an inventive engineer, it was always about numbers. Growing up, he often told my brothers and I that numbers were our “little friends”; however, it took me a while to realise that was not how numbers were generally perceived by others. Perhaps the most curious aspect of my father’s seemingly odd habit was that he was never pleased with just the answer, as he would often inquire about the ways in which my brothers and I had arrived at a particular answer, asking us to think carefully and consider alternative scenarios while furnishing the final response. Many years later, while working on my engineering degree and during my later graduate work in mathematics and statistics, I observed that some of my professors used the same strategies to encourage the careful overseeing of the ongoing thinking processes, particularly with respect to mathematical problem-solving.

As a novice high school teacher, I was under the faulty impression that most mathematics students are good judges of their own task performance outcomes. My repeated observations of students’ dismay and disappointment in getting poor results for tests for which they assumed they were well prepared proved me wrong. I noticed that in a few cases, the mismatch between the expected and achieved performance outcomes motivated students to seek ways to improve their task performance. In most cases, however, students' erroneous performance judgements seemed to cause them psychological frustration and despair, which had negative consequences for their academic (i.e., school related) achievement, subject disposition, and motivation. I also noticed that this phenomenon was widespread, affecting students of different ages and achievement levels in mathematics. Over the years of my teaching experience, this seemingly ever-present discrepancy between the students' expected and achieved grades created in me a deep desire to understand the nature of students' performance judgements. For instance, I wondered about the ways in which students judged their own procedural fluency, strategic competence, and adaptive reasoning in mathematics, as well as how they decided what to study and when to stop practising a new topic. Curiously, students did not seem to gradually become more accurate in their self-assessments, despite repeated episodes of testing and feedback. For instance, a student who repeatedly performed poorly on chapter tests would continue to reassure me of the remarkable job he had done on his latest test, only to be disappointed again by the poor results of his test.

Students' judgements about their test performance are an aspect of their metacognition known as *metacognitive knowledge monitoring* or *monitoring* for short. I based my study on a classroom intervention that I designed to explore the change and development of students' monitoring judgements called *calibration*. Calibration refers to the accuracy of people's monitoring judgements about their own task performance; to the extent that a person's monitoring judgements are accurate, the person is considered to be calibrated. Students possibly make monitoring judgements on their own while thinking thoughts such as follows: "This chapter is hard to understand. I should read slower and take some notes."

The accuracy of students' monitoring judgements may affect not only students' learning and achievement, but also their subject-matter disposition and motivation. Despite decades of research, the question of whether calibration accuracy can be developed in a classroom context remains elusive. In this study, I sought to provide—at least partially—an answer for this question.

In this section, I outline the organisation of this dissertation and its five constituent chapters. Within this dissertation, I italicised important phrases and definitions, book titles, and significant terminology for emphasis. All students' names used in this thesis are pseudonyms. These pseudonyms were selected to depict the diversity of the group and do not reflect any specific cultural background. In Chapter One, I attend to the subtleties of the problem of miscalibration and explain the importance of the accuracy of students' monitoring judgements. Further in Chapter One, I highlight the purpose and significance of this study and introduce my research question. In Chapter Two, I focus on the historical roots and conceptualisation of the concept of metacognition, outline the theoretical framework of this study, and highlight relevant perspectives from mathematics education and other disciplines. I discuss the problem of conceptualisation of calibration and review the classroom interventions designed to improve calibration accuracy. I also highlight the findings, gaps, and contradictory results in calibration research and relate the rationale for this study. In Chapter Three, I examine the conceptualisation and the utility of the mixed methods methodology applied in this study, introduce the research participants and context, and explicate the research design and the intervention components. I address research quality considerations and explain the assumptions, limitations, and delimitations of my research design. I conclude Chapter Three with some comments on ethical views and perceived benefits of my study for mathematics students. In Chapter Four, I explain my quantitative data analysis procedures, describe my three rounds of the blended data analyses, and summarise the findings of this study. Finally, in Chapter Five, I discuss my findings in the context of proposing an answer for my research question. I further explain the educational implications of my study, describe its contributions, and explain the limitations of my findings. Moreover, in Chapter Five, I offer my perspective on the future of calibration research and share my final reflections, with an emphasis on placing my study within the discourse of the contemporary research on calibration and metacognition in mathematics education.

Throughout this dissertation, I use the phrases ‘performance outcomes’ and ‘performance outcome scores’ interchangeably and consider performance outcomes on curriculum-based classroom quizzes, tests, and exams as task performance outcomes. The phrase ‘academic achievement’ has been used to refer to students’ task performance outcomes such as their grades on non-standardised curriculum-based tests and exams.

Moreover, in alignment with contemporary calibration research, the terms ‘low-achieving’ and ‘high-achieving’ have been used as abbreviations for ‘academically low-achieving students’ and ‘academically high-achieving students’ to refer to students who achieve low- or high-performance outcome scores in mathematics in classroom settings, respectively.

In this study, the emphasis is exclusively on developing students’ calibration rather than on improving their performance outcomes (i.e., grades in the course). However, for the purpose of quantitative analysis, calibration is quantified by using students’ performance outcomes; therefore, the narrative in this dissertation inevitably includes references to students’ performance outcomes in relation to their calibration.

Statement of the Problem

The world is changing, and the future societies that students will participate in will be different from those they live in today. The only certainty seems to be the uncertainty of what the future will bring. In this constant state of flux, in order to fully participate in "an increasingly technological world" (Schoenfeld, 1992, p. 335) and have a fair chance at academic and professional opportunities, students should be flexible and innovative thinkers who possess adaptive thinking and problem-solving skills.

Concerning mathematics courses at the high-school level, good performance is important due to the mandatory status and gatekeeping (Moses & Cobb, 2001) function of the subject. As such, a high-quality mathematics education for students is deemed enticing. Further, success in high-stakes examinations such as Ontario's Education Quality and Accountability Office (EQAO) tests and in Alberta's grade 12 diploma examinations, seems to be on the wish list of not only students but also their parents, teachers, boards of education, and society at large. Moreover, the responsibility of adequately preparing students rests with many individuals, not the least of whom is the student. As Zimmerman (2002) noted, sharing the responsibility of students being adequately prepared implies that students should be able to judge their subject-matter preparedness and adjust their study behaviour accordingly. Unfortunately, it is not always easy for students to answer questions such as the following:

- Do I understand the question that I am being asked?
- Are the steps I am taking toward a solution appropriate and in the right direction?
- Am I cognitively prepared for my final exam?
- What does it take to be successful at school?
- Where should I focus my study efforts?

- Can I distinguish between what I know and do not know at this time?

Students' capability to accurately judge their academic task performance relies on their metacognition. Scholars (e.g., Desoete, 2009; Lucangeli et al., 1998) have suggested that metacognitive skills are related to students' academic task performance in general and to mathematical task performance in particular. Regardless, researchers have concurred for decades that students—ranging from the very young to advanced university students—often lack the capability to accurately judge their own task performance (e.g., Bol & Hacker, 2001; Dunlosky & Lipko, 2007; Klassen, 2002; Pajares & Kranzler, 1995; Pashler et al., 2008; Rutherford, 2017a; Stoten, 2019) in academic domains such as school mathematics. In other words, students at various levels of education experience an illusion of knowing (Glenberg & Epstein, 1985; Kruger & Dunning, 1999), whereby they think they know the material, but they do not. This discrepancy may substantially affect their achievement and motivation. What makes this problem more pronounced is that the least-skilled students are the least aware of their erroneousness, leading to the coining of the phrase 'unskilled but unaware' by Kruger and Dunning (1999). Researchers (e.g., Alexander, 2013; Bol & Hacker, 2001; Brown, 1978; Dunning et al., 2003; Kruger & Mueller, 2002; Tobias & Everson, 1996, 2000; Zimmerman, 1986) have suggested that unskilled (i.e., academically low-achieving) students are often surprised and dismayed by how poorly they performed on a test for which they believed they were well prepared. Others (Hacker et al., 2008a) have found that learners who can accurately judge their performance are better able to manage their time, intensify or redirect their attention, and utilise other cognitive resources to complete academic tasks.

Some researchers (e.g., Alexander, 2013; Dunlosky & Thiede, 2013; Horgan, 1990) have suggested that the degree of the accuracy of students' monitoring judgements, at least in part, motivates students and guides their study behaviour. In other words, as noted in research (Lin & Zabrucky, 1998; Nietfeld et al., 2005; Stone, 2000; Wiley et al., 2005; Winne, 1995, 1996), students may decide not to spend time and effort studying unless they are aware of the existence of a misalignment between their estimations of task performance outcomes and their actual task performance outcomes. This occurrence underlines the importance of the accuracy of performance judgements (Schraw et al., 1993), which is also called calibration. While better calibration is linked to improved task performance in general (Butler & Winne, 1995; Chen, 2002; Schraw et al., 1993), researchers (e.g., Bellon et al., 2019; Cardelle-Elawar, 1992, 1995; Desoete & Roeyers, 2006; Stillman & Mevarech, 2010; Zhao et al., 2011) have suggested that metacognitive strategies, including calibration, are critical to improving learning and task performance in mathematics. An inspiration for this study was that some researchers have suggested that calibration is malleable (Schraw et al., 1993) and can be enhanced in classroom contexts (Huff & Nietfeld, 2009; Labuhn et al., 2010) using carefully designed pedagogical interventions (Rutherford, 2017a) in academic fields such as high school mathematics.

An Overview, the Purpose, and the Research Question of This Study

The purpose of my study was to explore the development and the potential change in students' calibration over time, in the context of a pedagogical intervention. In Chapter Three, I describe my research design in detail. However, in the next section, I provide a brief overview of how I implemented my classroom intervention.

My study of students' calibration included the three phases of pre-intervention, intervention, and post-intervention. I was the classroom teacher in a year 11 mathematics course and there were 24 students in the course. Participation was on a voluntary basis and the students were informed that they could stay in the course without participating in research procedures.

Since students' signed consent and assent forms were directly delivered to the main office at the host school and not to me, I did not know who the actual research participants were during the semester.

Within the pre-intervention phase, I implemented an open-ended questionnaire to collect data on students' awareness of their own calibration. During the intervention phase, after writing each weekly test, students made a performance estimation of their expected grades. In my attempt to support students' monitoring skills, students had the opportunity to choose their own test questions from a list of potential test questions that I made available to them. When I returned their graded tests, I encouraged the students to notice the gap between their estimated and achieved performance outcomes if such a gap existed, graph the estimated and achieved performance outcomes on their calibration graph templates, and complete a personal reflection notecard.

The testing, graphing, and reflection episodes were repeated every week between February 8, 2016 and May 13, 2016, except for the week of March Break. The last instance of testing and estimation took place when students wrote their final exam on June 14, 2016. The intervention phase included my explicit monitoring strategy training for students in two ways. First, during the intervention phase, each Tuesday, I started the daily classroom activities by teaching a session that I called a *calibration lesson*. Each lesson was about 15 minutes long during which I introduced calibration-related concepts. Second, I started a classroom routine in which, during the daily problem-solving group work, I approached the students with my *What*, *Why*, *How*, and *Can you Check and Reflect* questions. To explain, in order to encourage the students to monitor their thinking and problem-solving processes, I asked the students to be ready to explain *what* problem-solving method they were using, *why* they thought it was the right method, *how* well the method worked, and whether the solutions were accurate and meaningful.

My post-intervention phase included two steps. First, I implemented another open-ended questionnaire inquiring about students' awareness of their own calibration. Second, I conducted the follow-up interviews after the end of the school year, in the first two weeks of July.

As the classroom teacher, I had access to students' performance outcomes. However, due to the research ethics considerations, throughout the study, I did not collect or keep any record of students' performance estimations. A copy of students' classroom tests was sent to the main office at the host school to be kept in the vault as my research data, and the original tests were returned to the students. Moreover, students' questionnaires, reflection notecards, and calibration graphs were collected in large envelopes that I sent around the classroom and were sent to the vault. Once a week, a school office administrator brought back the envelope containing students' calibration graphs so that the graphs could be updated by the students. I accessed all my research data after the interviews were completed. This was two weeks after the end of the school year.

In conducting this study, the research question that I pursued was as follow: *In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom?* This research question reflects the centrality of the concept of *change* in students' calibration to my study; a notion which I pursued both theoretically by exploring the change in students' calibration awareness over time and practically by examining whether the students' calibration accuracy improved over time.

Significance of This Study

No one is perfectly immune to flawed self-evaluation. In an ideal world, people would always be able to provide a clear and accurate evaluation of their own competencies in different areas, including academics. As Miller and Geraci (2011) noted, however, errors in self-evaluation are routinely made by doctors, nurses, managers, accountants, and other trusted professionals. Thus, students' misjudgements about their task performance should not be surprising. The accuracy of students' task performance estimations, also called calibration is an aspect of their metacognitive monitoring, and the necessity of teaching students metacognitive strategies has widely been recommended in research in the past four decades (e.g., Alexander, 2013; Bateson, 1972; Brown, 1994; Desoete, 2008; Flavell, 1979; Garofalo & Lester, 1985; McCormick, 2003; Nietfeld & Schraw 2002; Polya, 1973; Rutherford, 2017a; Schoenfeld, 1985, 1987b, 1992; Thomas, 2003, 2006, 2013; Tobias & Everson, 1996, 2002). Bandura (1986) posited that "what people think, believe, and feel affects how they behave" (p. 35). In a broad sense, in this study, I focused on what students thought of their task performance, what they were confident about, and what they thought might affect their performance judgements, with the goal of developing students' calibration.

As a mathematics teacher, it has been my experience that teaching students the mathematical concepts, methods, and procedures may not always be enough for a meaningful learning experience and for successful task performance. Other skills such as metacognitive skills are required to ensure engagement (Cunha & Heckman, 2006) and to afford students the sensibility to monitor their thinking processes while on academic tasks. This sensibility is possibly not always easy for students to attain on their own, as it requires the knowledge of one's cognition for the deliberate reflection on tasks, as well as self-directed regulation of one's thinking processes.

However, not all the tasks ahead of students are school-related tasks. As students engage in life-long novel learning situations, those students who attempt to prepare themselves for higher education, joining the workforce, or both, are uniquely situated to benefit from classroom interventions aimed at developing their calibration, significant to realising success in both academic and non-academic tasks (Brown, 1978; Flavell, 1976, 1979; Veenman et al., 2004).

Researchers (e.g., Glenberg et al., 1987; Nietfeld et al., 2005; Schraw et al., 1993) have demonstrated the importance of optimal calibration accuracy for sustained effort and the efficient allocation of attention, and have attested to the role that calibration plays in students' awareness of their own academic strengths and weaknesses. By focusing on the development of students' calibration in a mathematics classroom in this study, I provided opportunities for students to practice, communicate, reflect upon, and adjust their monitoring strategies and performance judgements. As agents of their own learning and achievement, students were encouraged to set challenging and yet achievable task performance goals and regulate thinking activities to ensure successful task performance. Schoenfeld (1987b) proposes that the most important contribution of metacognition to the learning of mathematics rests on students' knowledge attainment about their own cognition and their development of adequate monitoring strategies. Thomas, Anderson, and Nashon (2007) suggest that metacognition inspires optimism by providing an alternative to deficit theories of cognition that so often portray students' learning potential as fixed and beyond the reach of any form of pedagogical intervention. Further, Hattie (2013) posits that the accuracy of students' judgements should be sufficiently taken into account in considering how students learn best, how to help them construct knowledge, and how to ensure that a greater number of students learn more. These comments in part, underline the significance of studies that aim to develop students' calibration, such as mine, in supporting students' lifelong journey of learning and achievement.

Chapter One Summary

In this introductory chapter, I introduced the concept of calibration as the extent to which students' monitoring judgments about their task performance outcomes aligns with their actual task performance outcomes and discussed the importance of calibration accuracy for students' learning, motivation, and achievement. I also explained the organisation of this dissertation and what the five chapters included in this document entail. Further, I stated the problem which motivated my study and elaborated on the importance of teaching mathematics students metacognitive monitoring strategies which may support their learning and achievement. In Chapter One, I also explained the purpose and significance of my study, and identified my research question as follows: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom?

In Chapter Two, I present a detailed review of the research literature on several aspects of metacognition, including its historical roots, conceptualisation, and constituent components in the context of Nelson and Narens's (1990) theory of metacognition, which is the theoretical framework for my study. I also delineate the social dimension of metacognition, situate my study within Bandura's (1986) social cognitive theory, and discuss the perspectives on metacognition from mathematics education and other disciplines. Moreover, I discuss the conceptualisation, assessment, and the educational importance of calibration, and review the classroom interventions that aimed to improve students' calibration. Finally, in Chapter Two, I outline calibration research findings, gaps, and inconclusive research results, explain the rationale for my study, and provide a chapter summary.

Chapter Two

Review of Literature

Introduction

In this chapter, I review the concept of metacognition and its constituent components—*metacognitive knowledge*, *metacognitive knowledge monitoring*, and *metacognitive control*—with a focus on the notion of the accuracy of metacognitive knowledge monitoring judgements, also called calibration. I use the term metacognition, as Flavell (1979) defined, to mean one's ability to monitor, evaluate, and plan one's thinking processes. Valdez (2013) cited that the phrase metacognitive knowledge monitoring, or monitoring for brief, pertains to "learners' knowledge of strategies that support cognition, as well as knowledge of conditions that dictate how to implement useful strategies to enhance content learning" (p. 141). The term calibration is often used to refer to a measure of the accuracy of monitoring judgements and is defined as the degree to which learners' perceptions of task performance, including school subject test performance, correspond with the objective evaluations of their actual task performance (Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; Nietfeld, Cao, & Osborne, 2006a).

The choice of the term calibration may sound strange to some readers. I sympathise, as without finding a convincing answer, I have also wondered why calibration, which is a technical term in engineering and instrumentation, is used to refer to the accuracy of people's monitoring judgements. I am also curiously reminded of some other unusual terminology such as *triangulation*, a navigation and land-surveying term, which researchers use to refer to the cross-verification of the results from a variety of data sources.

Setting the terminology aside, a discussion on calibration is meaningful only against the backdrop of metacognition. In this chapter, I explore metacognition and its historical roots and conceptualisation, followed by a review of some metacognition theories pertinent to this study. I then introduce and substantiate Nelson and Narens's (1990) theory, which is "now a classic theory of metacognition" (van Overschelde, 2008), as the theoretical framework of this study. Subsequently, I explore the social aspects of metacognition within the social cognitive theory of Bandura (1986) and introduce calibration as the accuracy of monitoring judgements, highlighting pertinent issues in conceptualisation and assessment of calibration. Further, I review the classroom studies aimed at developing students' calibration and elaborate on pertinent findings, gaps, and contradictory results in calibration research. I conclude Chapter Two by presenting the rationale for this study and a chapter summary.

Metacognition: Historical Roots and Conceptualisation

Metacognition, which Tobias and Everson (2002) identified as the most intensively studied cognitive process in developmental psychology, is defined as one's ability to monitor, evaluate, and plan one's thinking processes (Flavell, 1979). This is the definition that I used in this thesis. The commonly used conceptualisation of metacognition as thinking about thinking disguises a much more complex cognitive phenomenon that has baffled scientists, philosophers, and educators for centuries. To explain, the discussion concerning what is generally known as *Comte's Paradox* was undertaken during the 18th and 19th centuries by Auguste Comte (1798–1857), the French founder of positivism, who argued that *human introspection* (i.e., monitoring) is impossible and paradoxical because the observing and the observed organs are one and the same. It took science more than a century to demonstrate that Comte's Paradox was not a paradox at all; as Dunlosky and Metcalfe (2009) explained, "For introspection to happen, just a portion of the human brain was needed to look back upon itself" (p. 12). Today, despite numerous attempts to define metacognition, an agreement on how best to delimit the construct remains elusive (Hacker, Keener, & Kircher, 2009; Tobias & Everson, 2000; Veenman, Van Hout-Wolters, & Afflerbach, 2006). Nevertheless, as Hacker et al. (2009) noted, the immensity of the global research literature on metacognition over a span of five decades suggests that the disagreement on conceptualisation has not deterred keen investigators.

From Piaget to Flavell

The origins of the study of metacognition as a construct separate from cognition can possibly be traced back to Hart's PhD dissertation in 1965 (Metcalfe & Shimamura, 1994) focusing on Feelings of Knowing (FOK). That is, when one is asked a question pertaining to semantic knowledge such as "Which city is the capital of Nigeria?", people can usually quickly and effectively indicate whether they know the answer, although they may not make an attempt to retrieve the relevant information. Nevertheless, the official frame into which Hart's doctoral dissertation ideas fit was brought forward by Wellman and Flavell in the late 1970s almost a decade later, categorising FOK as metacognitive judgements. As Dunlosky and Metcalfe (2009) noted, although the rise of the Metacognition School of Psychology can predominantly be attributed to the efforts of Flavell and his colleagues, notably Wellman, the birth of this approach can be traced back to Flavell's assimilation of Jean Piaget's child development theory (Flavell, 1963) in his seminal book, *The Developmental Psychology of Jean Piaget.* Possibly, Flavell's notion of metacognition stemmed from Piaget and his colleagues' theoretical work on children having thoughts about their own thinking, which Flavell (1963) described as "the crowning achievement of intellectual development, the final equilibrium state toward which intellectual evolution has been moving since infancy" (p. 201). Some researchers (e.g., de Bruin & van Gog, 2012; Dunlosky & Metcalfe, 2009) have proposed that Piaget's approach has been directly responsible for Flavell's theorising about the importance of thinking about thinking. For Piaget, metacognitive thought—that is, the thought that can be directed and controlled by the thinker—is conscious, intelligent, and communicable. In Piaget's theories, metacognition seems to be completely absent before the ages of 11 or 12 years, since it has been theorised (e.g., Fox and Riconscente, 2008) that the arrival at metacognitive thought requires the transformation of children's social and intellectual egocentrism into adults' relativistic and socialised thought.

Elaborating on the historical development of metacognition, Schneider and Lockl (2008) explain that early research focused on the knowledge about memory or, as coined by Flavell (1979), *metamemory.* Flavell and Wellman (1977) conceptualised metacognitive knowledge as metamemory and offered a taxonomy for metamemory, distinguishing between the sensitivity category (i.e., procedural metacognitive knowledge) and the variables category (i.e., declarative metacognitive knowledge). Schneider and Lockl (2008) thus theorise on the prominence of metamemory in Flavell's work, which was focused on metacognitive knowledge, leaving little room for monitoring and control processes. In my review of Flavell's work, I could not locate a 'system' or 'model' to explain how control processes are influenced by the monitoring processes, which is the focus of my study. This connection was later made and articulated by Nelson and Narens's (1990) theory of metacognition, which is why I chose Nelson and Narens's (1990) theory as the theoretical framework for this study.

From Flavell to Nelson and Narens

Inspired by Piaget's work, Flavell (1976, 1979) coined the term 'metacognition' to describe any cognitive activity about cognition. In the years that followed, as van Velzen (2016) elaborated, "Flavell would study metacognition mainly as children's metamemory or their understanding of their own memory" (p. 14). De Bruin and van Gog (2012) opined that although Flavell (1979) was particularly successful in theorising about metamemory (i.e., people's knowledge of the content of their memory), it was not until Nelson and Narens's (1990) seminal work that a general theory of metacognition—a system to explain how metacognitive processes acted upon and interrelated with cognitive processes—was presented.

This thesis is the study of the development of students' monitoring judgments, or calibration, and the relevance of Nelson and Narens's (1990) theory for this thesis lies in its utility in demonstrating that effective control depends on accurate monitoring. In other words, good monitoring presumably gives students a chance at employing appropriate strategies such as allocating time, attention, and effort to achieve their set goals. In Stoten's (2019) view, metacognition facilitates students' awareness of how to regulate their learning, thus playing a vital role in students' management of their own academic journeys. Further, as Tobias and Everson (2002) argued, "learners who accurately differentiate between what has been learned previously and what they have to learn are better able to focus attention and other cognitive resources on the material to be learned" (p. 1).

The Conceptualisation and Constituent Components of Metacognition

The ability to reflect upon and regulate our thoughts and behaviour are considered by some (e.g., Bandura, 1986; Metcalfe & Shimamura, 1994) to be at the heart of what makes us distinctively human. Despite the growing evidence that identifies metacognition as a vital component of intelligence and cognition, an agreed-upon conceptualisation of the term metacognition remains elusive. In my search for a clear working definition to operationalise metacognition, I realised that within the discipline of psychology, Flavell (1976) authored one of the earliest definitions of metacognition:

> Metacognition refers to one's knowledge concerning one's own cognitive processes and products, or anything related to them….Metacognition refers, among other things, to the

active monitoring and consequent regulation and orchestration of these processes in relation to the cognitive objects or data on which they bear, usually in the service of some concrete goal or objective. (p. 232)

In 1985, Flavell offered a slightly different definition by describing metacognition as "any knowledge or cognitive activity that takes as its object, or regulates, any aspect of any cognitive enterprise… its core meaning is cognition about cognition" (p. 104). The latter definition is less technical than the former and closer to the commonly used notion of thinking about thinking.

Although metacognition is the frequently used term in research literature, other terms have also been used to refer to one's examination of one's own thoughts. For example, Piaget spoke of *reflective abstraction* (Dubinsky, 1991; Fox & Riconscente, 2008), with similarities to the notion of metacognition. Skemp (1979) used the phrase *reflective intelligence*, which he defined as "the ability to make one's own mental processes the object of conscious observation" (p. 175), similar to the concept of metacognition.

Curiously, the earlier definitions were offered without any consensus on the main components of metacognition. For instance, Flavell's (1976) definition of metacognition encompasses not only metacognitive knowledge, but also monitoring and control processes and their relationship within the construct of metacognition. This definition has some implications for teachers, as it indicates that, possibly, students' control-related activities (e.g., deciding to re-read a word-problem if they did not understand it the first time) are a consequence of students' monitoring-related activities (e.g., while reading the word-problem, students may think whether or not they understood the question). Hence, this definition of metacognition implies that enhancing students' monitoring skills may support their task performance in school subjects.

In her review of literature on metacognition, Lai (2011) pointed out that after Flavell, many of the subsequent users of the term metacognition remained relatively faithful to Flavell's original conceptualisation. For instance, Cross and Paris (1988) described metacognition as "the knowledge and control that children have over their own thinking and learning" (p. 131). Further, Nelson and Narens (1990, 1994) and Narens (1996) defined metacognition as a state of cognition at a meta level representing cognition. Hennessey (1999) defined metacognition as the awareness of one's own thinking and the content of one's conceptions, an active monitoring of one's cognitive processes, and an attempt to regulate one's cognitive processes in relationship to further learning. According to Kuhn and Dean (2010), metacognition refers to the "awareness and management of one's own thoughts" (p. 270), Martinez (2006) described metacognition as "the monitoring and control of thought" (p. 696), and van Overschelde (2008) conceptualised metacognition as knowing about knowing. These varied definitions attest to the complex and multidimensional (Efklides, 2008) nature of metacognition.

Not only the conceptualisation, but also the main components of metacognition have been greatly debated and disputed without an overall agreement. Flavell (1979), Brown (1980), and Garner and Alexander (1989) referred to metacognition as the knowledge, awareness, and control of one's cognitive activities, while Paris and Winograd (1990) argued that metacognition only encompasses knowledge but not control or regulation. Following Brown (1978) and concurring with Flavell's (1979) conceptualisation of metacognition, some researchers (e.g., Baker, 1989; Cross & Paris, 1988; Jacobs & Paris, 1987; Schraw, 1997; Schraw & Dennison, 1994) have suggested two components for metacognition: knowledge of cognition and regulation of cognition.

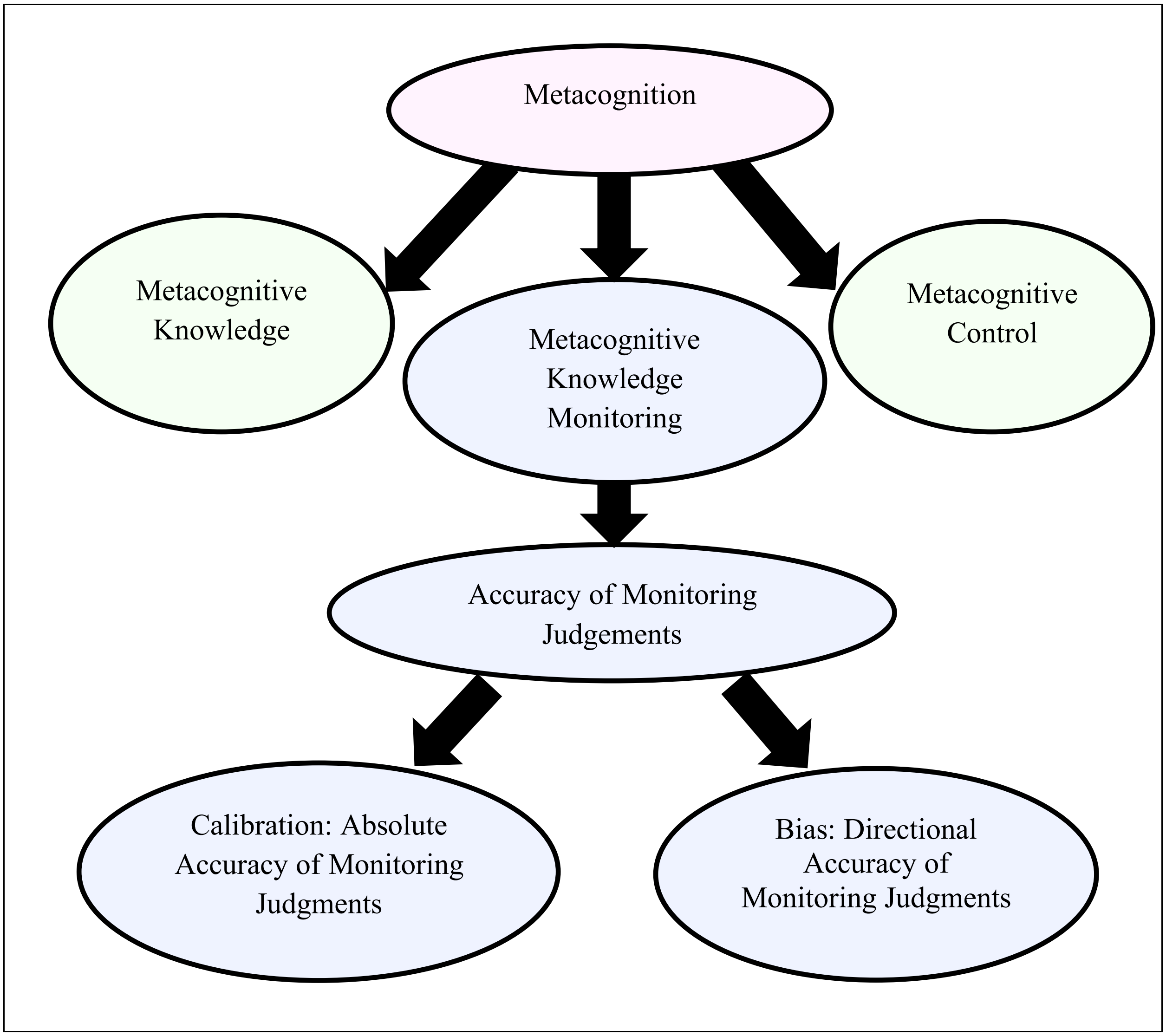
In the absence of a clear and definitive conceptualisation of the components of metacognition, the individual researcher's discretion on which scholarly discourse to subscribe to and which components to use in research prevails. In this thesis, I have adopted Flavell's (1979) conceptualisation of metacognition to mean one's capability to monitor, evaluate, and make plans for one's thinking processes.

Further, I have adopted the additional perspective that considers metacognitive knowledge, metacognitive knowledge monitoring (i.e., monitoring for short), and metacognitive control (Paris & Oka, 1989; Pintrich, Wolters, & Baxter, 2000; Schneider & Artelt, 2010; Winne & Hadwin, 1998) as the main components in conceptualisation of metacognition. These components are further classifications that provide important distinctions for understanding the primary role of the accuracy of students' monitoring processes, which are indicative of their self-evaluative judgments (Nietfeld et al., 2008b) in the invocation of students' control processes and are particularly well-suited to the purpose of this study. In the following section, I elaborate on the three constituent components of metacognition as used in this thesis.

To provide a conceptual context for introducing calibration meaningfully, I explored the construct of metacognition and its constituent components, metacognitive knowledge, metacognitive monitoring, and metacognitive control. Figure 2.1 represents a conceptual taxonomy of the interconnected concepts discussed in this thesis. To explain what the arrows communicate, in this study, following my review of the broad construct of metacognition, I attended to the three aforementioned components of metacognition. However, of these three components, I kept a focused lens on students' metacognitive monitoring judgments and the accuracy of these judgments both as a positive value, called calibration, and as a directional positive or negative value, called bias to explore students' overestimations or underestimations.

Figure 2.1

Taxonomy of the Interconnected Concepts Explored in This Study



Metacognitive Knowledge. Flavell (1979) described metacognitive knowledge as the knowledge about one's own cognitive strengths and limitations, including the knowledge about all the external or internal factors that may interfere with one's cognition. In this thesis, I adopted Flavell's definition of metacognitive knowledge as any attained knowledge that has to do with cognitive matters. Flavell classified such knowledge into the three categories of *person* *knowledge*, *task knowledge*, *and strategy knowledge*. In Flavell's view, person knowledge includes anything that a person may believe about the nature of oneself and others as cognitive beings. For example, a student may believe that she/he learns algebra easier than geometry or that her/his classmates are better at graphing than she/he is. In this study, helping students construct person knowledge about calibration was paramount. Students were encouraged to participate in calibration sessions and monitoring strategy training sessions and to reflect on their calibration to develop their person knowledge.

According to Flavell (1979), task knowledge includes the knowledge about the nature and demands of different tasks. For instance, a student may know that it takes her/him longer to complete a unit test than a quiz. Developing students' task knowledge was an important aspect of this study in which, as a pedagogical intervention, students had the opportunity to choose their own test questions. Finally, Flavell (1979) viewed strategy knowledge as the knowledge of strategies that are likely to be useful in completing a task such as knowing that drawing a diagram might be helpful in solving a three-dimensional problem in trigonometry. In this study, through monitoring strategy training, I aimed to improve students' strategy knowledge.

Research on metacognition provides a variety of different theoretical frameworks for the conceptualisation and categorisation of metacognitive knowledge. For instance, several researchers (e.g., Kuhn, 2016; Schraw, Crippen, & Hartley, 2006; Paris & Winograd, 1990; Schraw & Moshman, 1995; Kuhn & Dean, 2010) have used the concepts of *declarative* and *procedural* metacognitive knowledge to distinguish between what they perceive as different types of metacognitive knowledge. Kuhn and Dean (2010) described declarative metacognitive knowledge to be a students' general understanding about thinking and knowing, while Schraw et al. (2006) defined declarative metacognitive knowledge as the knowledge about oneself as a learner and about the factors that may influence one's performance. Paris and Winograd (1990) proposed that declarative metacognitive knowledge is the knowledge that students reflect upon when answering the question, "Do I know this?" Some researchers (e.g., Cross & Paris, 1988; Kuhn & Dean, 2010; Schraw et al., 2006) have described procedural metacognitive knowledge as the awareness and management of cognitive processes, including the knowledge about strategies. Procedural metacognitive knowledge generally answers the question, "Do I know how?" *Conditional metacognitive knowledge* was the third category, added by Paris and colleagues (1984), involving the knowledge of *why* and *when* to use certain strategies. Paris et al. hence suggested that as a broad category, metacognitive knowledge can be divided into three distinct areas of (a) declarative metacognitive knowledge, (b) procedural metacognitive knowledge, and (c) conditional metacognitive knowledge.

In a general sense, metacognitive knowledge pertains to the knowledge concerning the human mind, its capabilities, and its doing. Despite the differences in their approaches to learning and development, researchers agree (e.g. Pintrich & Schunk, 2002) that with development, students become more aware of their own thinking and more knowledgeable about cognition in general. This awareness has been referred to by several names such as metacognitive knowledge, metacognitive awareness, and self-awareness. Students use their metacognitive knowledge to *monitor* their ongoing thinking processes. This process is often called knowledge monitoring (Hartwig et al., 2012; Isaacson & Fujita, 2006; Pintrich et al. 2000; Tobias & Everson, 1996, 2000, 2002), or just monitoring, for brief. As such, the development of metacognitive knowledge in students has a prominent role in any study of students' metacognition, including this study.

To explain, as students come to know about the construct of calibration by for instance, partaking in calibration sessions and classroom conversations, they develop some knowledge about calibration as well as about their own capabilities and ways of becoming better calibrated than they already are. In the context of a mathematics classroom, students may develop the knowledge that they know several ways of solving a second-degree equation (declarative metacognitive knowledge), that graphing, factoring, completing the square, and using the quadratic formula are some of those ways (procedural metacognitive knowledge), and that it is easier to solve a second-degree equation by using the quadratic formula than by completing the square (conditional metacognitive knowledge).

According to Efklides (2008), implicit in the two classic definitions of metacognition put forward by Flavell (1979) and Nelson and Narens (1994) is the assumption that students have some metacognitive knowledge about the content, procedures, and demands of what they are monitoring and thus, metacognitive knowledge is foundational to monitoring. Nelson and colleagues (1998) suggested a broader conceptualisation of metacognitive knowledge by identifying metacognitive knowledge as the general knowledge a person has about cognitive tasks and processes. This conceptualisation of metacognitive knowledge includes self-knowledge, which may draw on the subjective feelings and experiences and cultural theories and beliefs a person maintains. I posit that any study of students' calibration, including my study, is ultimately the study of the development of students' monitoring of their declarative, procedural, and conditional metacognitive knowledge; therefore, in my study of calibration, I classified metacognitive knowledge into the categories of declarative, procedural, and conditional knowledge, which best suited the purpose of my study.

Nelson and colleagues (1998) theorised a social side to metacognition involving people's metacognitive assessments of others' knowledge. For example, a person may think, "What does my friend think, and why does she think this way?" Hence, Nelson et al. (1998) explained that a rich source of metacognitive knowledge is what people draw upon in their social interactions and communications, as well as to provide social support and to make group decisions. This conceptualisation of metacognitive knowledge seems to be much broader than the initial definition of Flavell (1979). Kuhn (2016) proposed that the epistemological beliefs about the nature, truth, and validation of knowledge as well as the means and methods of knowledge construction can also be considered as metacognitive knowledge.

Researchers (e.g., Pintrich & Schunk, 2002; Thiede et al., 2003; Wiley et al., 2005) suggest that the development of students' metacognitive knowledge is critical for students' monitoring of a variety of tasks. These findings also emphasise the importance of teaching such knowledge explicitly to students. Within the present study, the intervention components and classroom discussions were meant to continuously enrich and update students' metacognitive knowledge and to encourage students' self-awareness of their own cognitive strengths and limitations so that they might develop their monitoring strategies further and eventually become better calibrated on mathematical tasks. In addition to the importance of the development of metacognitive knowledge in students, Schraw (1994) suggested that various types of metacognitive knowledge influence each other. For instance, knowing that they are prone to making calculation errors (declarative metacognitive knowledge), students may consider double-checking their calculations (procedural metacognitive knowledge), while engaged in what they consider to be novel or complex tasks (conditional metacognitive knowledge).

Metacognitive Monitoring. The second component of metacognition is metacognitive monitoring, or monitoring for short, which has been described (Schraw & Moshman, 1995) as one's awareness of one's own task performance while engaged in a task. As noted earlier, monitoring is the short form of metacognitive knowledge monitoring (Hartwig et al., 2012; Isaacson & Fujita, 2006; Pintrich et al., 2000; Tobias & Everson, 1996, 2000, 2002), which indicates a close connection between monitoring processes and metacognitive knowledge in a sense that while monitoring is in progress, what is being monitored is metacognitive knowledge. Flavell (1979) suggested that metacognitive experiences that facilitate the monitoring of cognitive processes gradually enrich and refine metacognitive knowledge. In fact, according to Flavell, a reciprocal influence exists: not only the metacognitive insights experienced while monitoring cognition play a role in the development and refinement of metacognitive knowledge, but also, as empirically demonstrated by Schraw (1998), the enrichment of metacognitive knowledge appears to facilitate and enhance one's ability to monitor one's cognition. In Flavell's work, metacognitive monitoring is discussed in the context of metacognitive experiences. These experiences are the insights students gain while engaged in cognitive tasks. For instance, a student who thinks "I should draw a diagram. I am not sure I understand this question," undergoes a metacognitive experience.

While calibration has been linked to enhanced study habits (Gutierrez & Schraw, 2015; Horgan, 1990, Persky & Dinsmore, 2019; Stoten, 2019) and improved task performance (Bol, Hacker, O'Shea, & Allen, 2005; Butler & Winne, 1995; Chen, 2002; Rutherford, 2017a; Schraw et al., 1993, Stoten, 2019), researchers (e.g., Alexander, 2013; Bol & Hacker, 2001; Kruger & Dunning, 1999) suggest that for a variety of reasons, students are not very good at accurately judging their subject matter task performance against a perceived standard. In other words, they are not naturally very well-calibrated. Metacognition researchers (e.g., Hattie, 2013; Kruger & Dunning, 1999) suggest that, as counter-intuitive as it may sound, academically low-achieving students are often known to overestimate their own task performance, which possibly makes them less motivated to intensify or redirect their efforts to improve their task performance. While completing a novel or complex task, students presumably monitor their progress and allocate time, attention, and learning resources according to their own monitoring judgements (e.g., "I find this section difficult. I cannot stop now. I need more examples"). To some extent, whether students decide to continue or terminate their study session may depend on the decisions made by them based on their monitoring judgments while engaged in a task.

Monitoring judgements are classified into several categories. Schraw (2009) elaborated that at some point, while studying, students must decide if they have studied well-enough for successful task performance, unless they have simply run out of study time. Such judgement is called *Judgement of Learning* (JOL). For instance, a student may think to herself, "I think I have studied enough on this subject and will do well on my upcoming exam."

Another judgement occurs when students try to ensure they have answered the test questions adequately, called *Confidence Judgement*, also called *Judgement of Confidence* (JOC; Schraw, 2009). For instance, a student may think to himself, "I think I will get close to 80% on my test." For the purpose of this thesis, I studied the accuracy of students' JOCs also called calibration, the word used in this thesis, retrospectively (i.e., after completing a task). This is the recommended way to measure monitoring accuracy (Schraw, 2009) when the effectiveness of an intervention is under study. Jamieson-Noel and Winne (2003) explain that calibration is always reported as a positive number (i.e., an absolute value), indicating the discrepancy between estimated and achieved performance outcomes.

In my study, in addition to calibration, *bias* (Chen, 2002; Mengelkamp & Bannert, 2010) is also reported. The numerical value of bias can be positive if estimation is higher than the performance outcome or negative if estimation is lower than the performance outcome. Bias indicates students' overestimation or underestimation of their performance outcomes. According to Pieschl (2009), monitoring judgements can be categorised based on whether they are made before, during, or after task performance. Monitoring judgments that are made before task performance are called predictions or prospective judgements. Monitoring judgments made during task performance are called concurrent or online judgements and those made after task performance are noted as postdictions or retrospective judgements. Pieschl (2009) further asserts that metacognitive judgements can be classified based on granularity, either as local judgements specific to one item, or as global judgements across multiple items such as an entire test.

Four different types of prospective judgements have been used in monitoring research (Pieschl, 2009; Pintrich, 2000; Schraw, 2009; Winne, 2004): Judgements of Learning (JOLs), Feelings of Knowing (FOKs), Judgments of Confidence (JOCs) also called Confidence Judgments, and Ease of Learning (EOLs).

Schraw (2009) described JOLs as the judgements of subsequent recollection of recently studied material and FOKs as the judgements of subsequently recognising information that could not be recalled. Moreover, according to Schraw (2009), JOCs are the judgments of confidence about task performance and EOLs are the judgements about the relative ease of learning the material. Concurrent judgements made for performance evaluation while on a cognitive task are often divided into online confidence judgements and online performance accuracy judgements. Finally, Schraw (2009) described retrospective confidence judgements as the retrospective judgements of task performance. They are postdiction judgements of how well one has performed, made after completing all items on a task.

In this thesis, students' retrospective confidence judgements of their own classroom test performance were explored, several times. I decided to use students' retrospective judgments because calibration researchers (Crissman, 2006; Schraw, 2009; van Loon et al., 2013) consider these judgements to be more accurate than either prospective or concurrent judgements due to the availability of metacognitive knowledge (Flavell, 1979) in the categories of person, task, and strategy knowledge, which were discussed earlier in this chapter.

The retrospective judgements of task performance (Schraw 2009), also called judgements of confidence are considered to be indicative of monitoring processes (Hadwin & Webster, 2013; Stone, 2000) and are typically made when students are asked to rate their confidence (Stone, 2000) about their performance after completing an academic task. For instance, after writing a test, a student may rate his or her confidence at 85% which indicates an 85% confidence level in the overall accuracy of his or her test performance. Judgments of confidence or JOCs have been examined as both predictions (Nelson & Dunlosky, 1991) and postdictions (Dinsmore & Parkinson, 2013) at both global and local levels (Pieschl (2009).

Based on the classification suggested by Schraw (2009), Table 2.1 depicts the taxonomy of eight different monitoring judgements and provides the timing and a brief description of each judgment. The bold section in this table indicates the type of monitoring judgement that I used in my study.

Table 2.1

Taxonomy of Monitoring Judgements

Time of Judgement	Type of Judgement	Description
Prospective Judgements	Judgements of Learning (JOL)	Judgements of subsequent recollection of recently studied information
	Ease of Learning (EOL)	Judgements made prior to study about the relative ease of learning information
	Feeling of Knowing (FOK)	Judgements of subsequently recognising information that could not be recalled
Concurrent Judgements	Online confidence judgements	Judgements of confidence in one's performance
	Ease of solution judgements	Judgements about the accuracy of performance
	Online performance accuracy judgments	Judgements about the accuracy of performance
Retrospective Judgements	Ease of learning	Judgements after study or testing about the relative ease of learning information
	Retrospective judgements of confidence about performance accuracy (JOC)	**Judgements of how well one has performed after completing all test items**

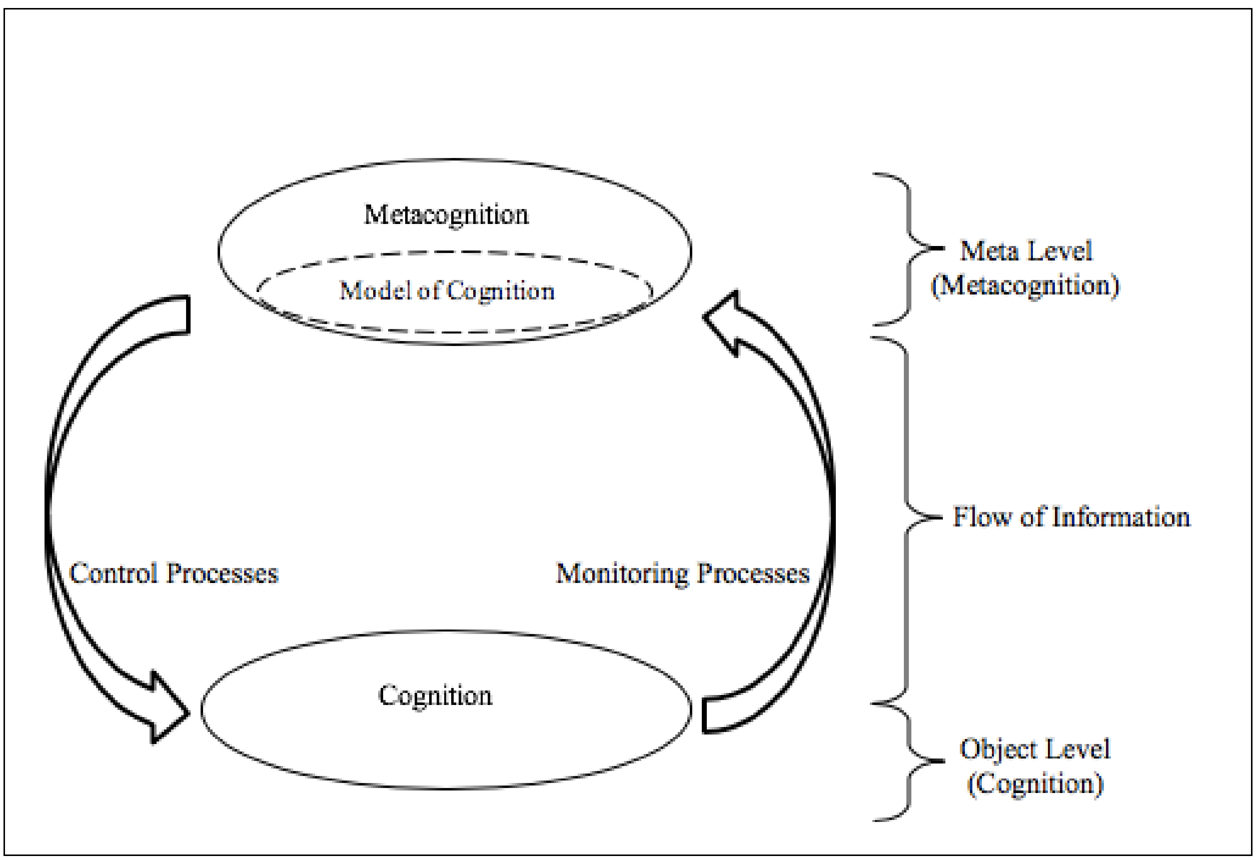
Metacognitive Control. In some theories of metacognition, including Nelson and Narens's (1990) theory of metacognition applied in this study, metacognitive control, or control for short, is identified as the third component of metacognition. Further, it is suggested that control processes are a result of monitoring processes. To explain, it is generally assumed that individuals are, at least to some extent, aware of their own beliefs or ideas about their own cognition. In other words, it is assumed (Flavell, 1979; Koriat, 2007; Nelson et al., 1998) that individuals possess some metacognitive knowledge and can report on their beliefs about their own cognition or reflect on them. Individuals then draw on this knowledge through self-monitoring, and form some kind of feeling or judgement (e.g., "I do not think that I know the trigonometric identities well enough to pass the upcoming test"), resulting in thinking about the conscious control of their cognition (e.g., "I think I should continue to practice with different models of trigonometric identities"). Researchers suggest that students' initial monitoring judgements, the subsequent control decisions regarding the use of attentional and cognitive strategies, and students' perseverance on cognitive tasks are closely interconnected (e.g., Papaleontiou-Louca, 2003; Pintrich & Schunk, 2002; Tobias & Everson, 2002). Within Nelson and Narens's (1990) theory of metacognition, control processes are conceptualised as the information that flows from the metacognition level to the cognition level, affecting the cognitive processes by initiating, continuing, or terminating an action. Hence, Nelson and Narens's theory separates metacognitive control from metacognitive knowledge. This is consistent with the view of those researchers (e.g., Pintrich et al., 2000; Winne & Hadwin, 1998) who have identified the general components of metacognition as (a) metacognitive knowledge, (b) metacognitive knowledge monitoring, and (c) metacognitive self-regulation and control, classifying the control processes as being driven by the monitoring processes; similar to this study.

Nelson and Narens's Theory: The Selected Theoretical Framework of This Study

Contemporary research on metacognition has often been guided by Nelson and Narens's (1990) theory of metacognition (Schraw et al., 2014), which, as noted by Hacker and colleagues (2009) has served as a versatile theory for the conceptualisation of metacognition as well as a heuristic for further theorising and empirical research. This theory conceptualises the overall processes of thinking as cognitive, such as using a formula to solve a problem, and metacognitive, such as monitoring the steps to ensure an accurate and meaningful solution. The theory further explains how learners use metacognitive processes to monitor and regulate their learning. The simplest version of the theory includes one object level (i.e., cognition level) and one meta level (i.e., metacognition level). In this theory, the information from the object level updates and informs the meta level through the monitoring processes, while the information from the meta level influences and controls the object level through the control processes. Figure 2.2 illustrates the elements included in Nelson and Narens's original theory of metacognition (Nelson & Narens, 1990) and the interrelations of these elements.

Figure 2.2

Nelson and Narens's (1990) Theoretical Framework for Metacognition



As depicted in Figure 2.2, Nelson and Narens (1990) based their theory on three underlying principles:

1. Mental processes are divided into "two or more" (Nelson & Narens, 1990, p. 125) interrelated levels. The basic structure, as illustrated in Figure 2.2, includes an object level and a meta level. This theory, however, can easily be extended to more than two levels (Nelson & Narens, 1994) for both monitoring and control processes. Van Overschelde (2008) explains that the object level is often associated with the thinking about an external object that is consistent with cognitive processes. For instance, "What I am seeing is a red car;" however, the meta level involves cognition about the object level, such as "Why am I thinking about that red car?", which is consistent with metacognitive processes.
2. The meta level contains a dynamic model of cognition, which also includes an individual's metacognitive knowledge. This principle concerns the way in which information flows within the model. The meta level acquires information and updates about the object level through the monitoring processes. It then sends information to the object level and hence influences the object level through the control processes.
3. There are two main processes: control and monitoring. Moreover, the monitoring processes informed by cognition continuously adjust the dynamic model of cognition within the meta level, which as stated by van Overschelde (2008), contains an individual's metacognitive knowledge. Hence, the theory explains the consequentiality of the accuracy of monitoring judgements for the selection and use of the necessary control processes (e.g., planning, on task monitoring, post hoc evaluation) that are necessary for successful studying and school task performance. As such, the pedagogical importance of

Nelson and Narens's (1990) theory of metacognition lies in demonstrating that teaching students effective monitoring strategies is necessary for the optimal selection and use of control processes.

The question of how metacognitive knowledge fits within Nelson and Narens's (1990) theory of metacognition deserves some elaboration. As Schneider (2008) noted, the conceptualisation of metacognitive knowledge has been a popular topic of research among developmental psychologists including Flavell and Wellman. In other areas, such as in the field of cognitive psychology, the scope of research has been much narrower. Cognitive psychology researchers (e.g., Metcalfe & Shimamura, 1994; Nelson, 1996; Nelson & Narens, 1990, 1994) have almost exclusively focused on procedural metacognitive knowledge. Particularly, Nelson and Narens (1990) used the terms 'information' and 'data,' instead of the term 'knowledge.' For instance, they defined monitoring as a data-driven dimension of metacognition (Nelson & Narens, 1994), which resonates with their exclusively quantitative methods in assessment of metacognition.

Elaborating on the monitoring and control processes within Nelson and Narens's (1990) theory, van Overschelde (2008) explains that the meta level contains (a) a dynamic model of the current state of the object level (Nelson & Narens, 1990, 1994), (b) a meta level goal state for the object level, and (c) the knowledge and strategies about the ways in which the meta level can influence the object level (Nelson, 1996) so that the object level may attain its meta level's goal. In this interpretation, metacognitive knowledge has a broad meaning, as van Overschelde (2008) refers to the knowledge that students possess about how the mind works and how its various functions can be controlled. Such metacognitive knowledge is explicit and factual, and it may affect all the judgment and control decisions that students make.

Further, van Overschelde (2008) argues that Nelson and Narens (1990, 1994; Nelson, 1996) have explicitly stated that the meta level contains the following three items:

1. A dynamic model of the current state of the object level based on the input from monitoring processes
2. A representation of a goal or goal state
3. A list of the known, possible control actions by which the meta level can influence or control the object level, details about when to use each control action, and the potential consequences

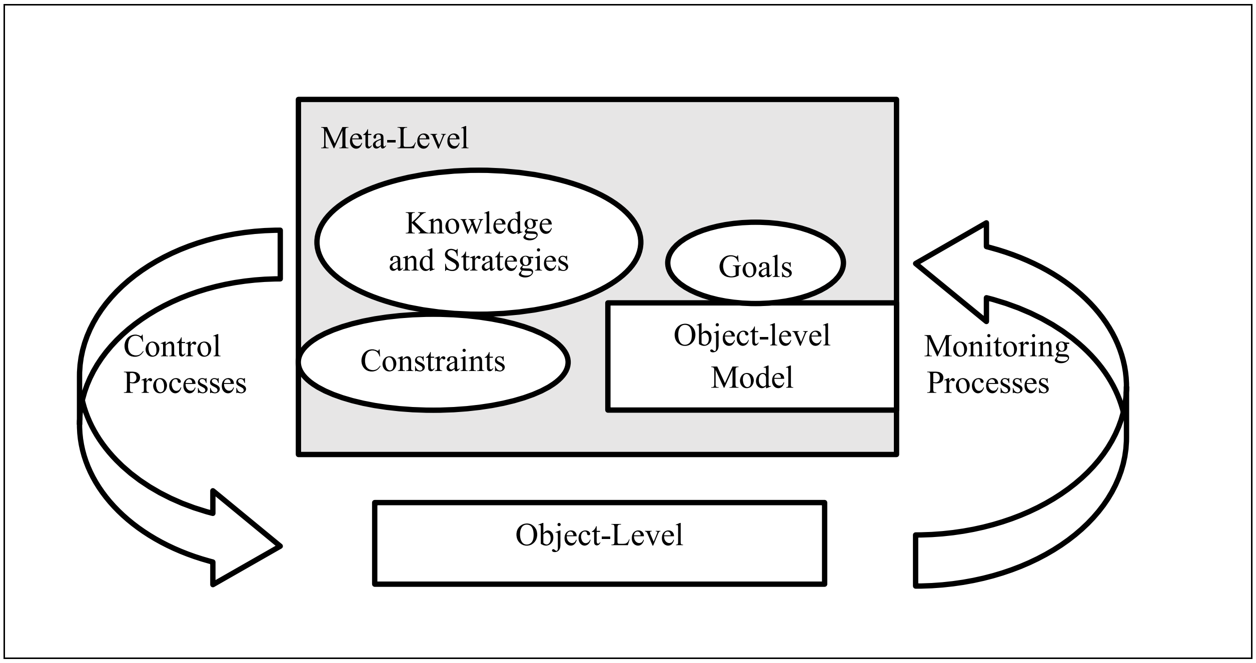
Further, van Overschelde (2008) suggests that Nelson and Narens's writings on the subject (1990, 1994; Nelson, 1996) imply that the meta level contains the following two items:

1. A list of the perceived constraints on potential control actions such as time limits, beliefs, and expectations
2. A judgement or decision-making process that evaluates the object level model and makes a decision about selecting the appropriate response or course of action to attain the intended meta level goal.

Figure 2.3 depicts van Overschelde's (2008, p. 50) visualisation of Nelson and Narens's (1990) theory of metacognition by demonstrating how metacognitive knowledge, which drives the five aforementioned components of the meta level, fits within Nelson and Narens's theory of metacognition.

Figure 2.3

A Representation of Nelson and Narens's (1990; 1994) Theory (van Overschelde, 2008)



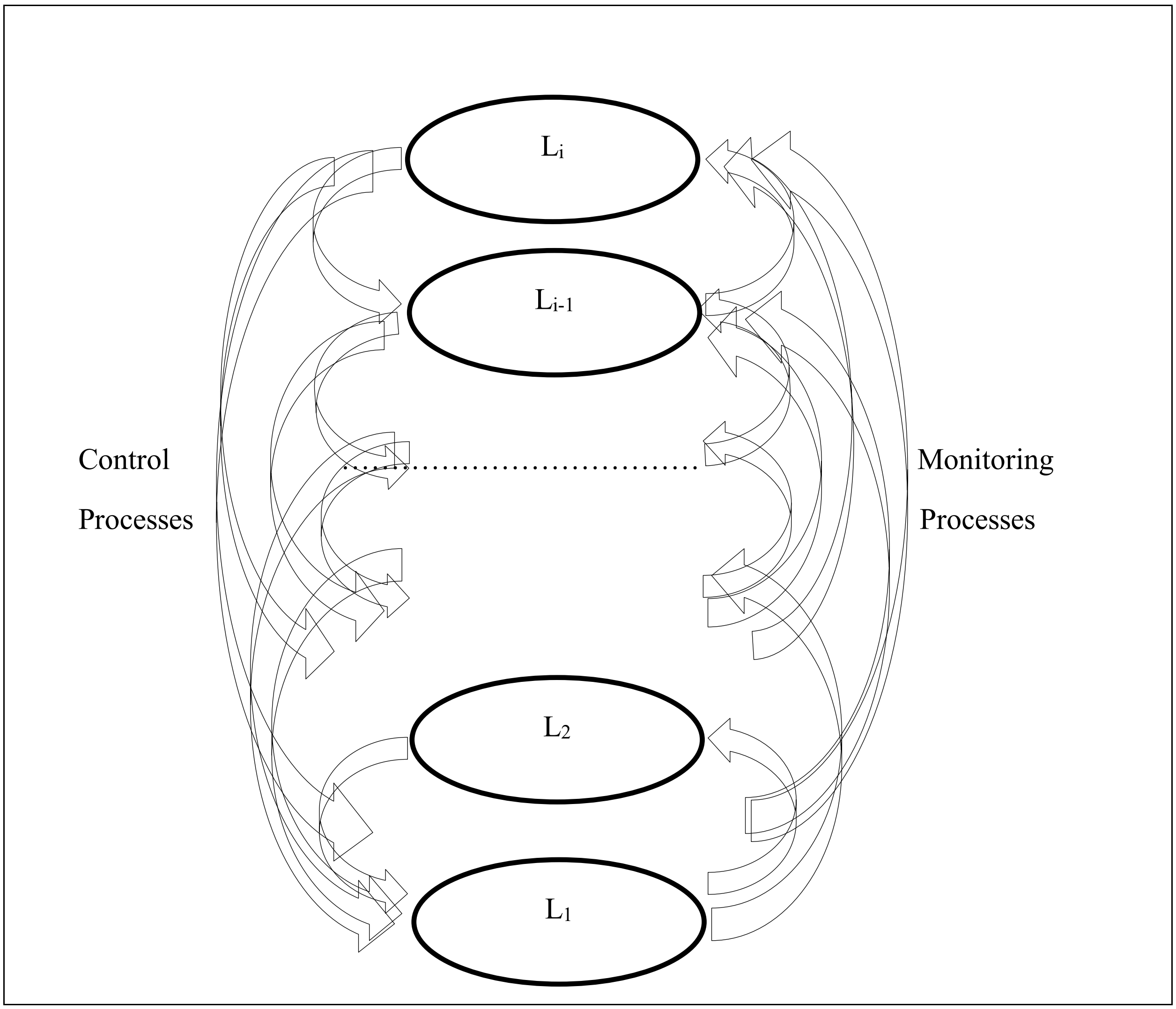
Nelson and Narens (1990) proposed that cognitive processes are split into two or more specifically interrelated levels, thereby allowing for the theory to be extended beyond the two basic levels. Later, Nelson and Narens (1994) extended the original model to include several interrelated levels using a similar development for the monitoring and control processes.

To explain, in this extended version, the meta level uses the information about the object level through the monitoring processes, and thus the model of cognition within the meta level is updated (i.e., the metacognitive knowledge within the meta level is updated). Nelson and Narens (1994), however, explained that this new state of the meta level might in turn act as the object level for the next level and this multilevel processing might be extended to a finite number of levels L_0, L_1, L_2, …, L_{i-1}, L_i, where L_i processes the information from its previous level L_{i-1} and perhaps from all other interrelated lower levels for which L_i is the meta level.

Figure 2.4 illustrates my understanding of the extended version (Nelson & Narens, 1994) of the original theory (Nelson & Narens, 1990) of metacognition.

Figure 2.4

A Multi-Level Extension of Nelson and Narens's (1990) Theory



Nelson and Narens (1994) thus explained that the critical issue does not concern the number of different levels of cognition and metacognition, but rather the relational aspects, such as which processes influence the others within the monitoring and control systems.

To provide an example from a mathematics student's perspective, suppose that after some reading (object level/cognition), Leila (a pseudonym) thinks, "I am reading this chapter too fast and may not remember much of it later. Perhaps I should slow down and take some notes while reading." This thought provides a monitoring input as Leila realizes that she is reading too fast. This understanding may potentially adjust Leila's model of cognition by adjusting her metacognitive knowledge (van Overschelde, 2008) at the meta level, which in turn informs the control processes. Subsequently, Leila may decide to slow down and take notes. Here, it is the input from the cognition at the object level that informs the meta level through the monitoring processes, and it is the metacognitive input that regulates the study behaviour through control processes. Based on Nelson and Narens's (1990) theory, without sufficient accurate monitoring processes, control processes might not be triggered. For instance, unless Leila suspects a problem with her study behaviour, she may neither decide to slow down nor to take notes. To emphasise, in this model, control processes are informed only by the meta level without receiving any information from the object level. In other words, the object level has no model of metacognition and might not be influenced by it. This is important because through monitoring processes, students use their knowledge about their cognition to evaluate their progress towards a learning or performance goal that they maintain at the meta level. Their evaluations update the model of cognition at the meta level, which contains the current state of their cognition at any time. Unless students' model of cognition at the meta level is constantly updated through their monitoring processes, they might not be metacognitively aware to make decisions to use relevant control processes to adjust their study behaviour towards achieving their set goals. This aspect partly explains why even a well-designed and well-implemented intervention to improve students' metacognition might ultimately fail, unless students' monitoring processes are enhanced through explicit monitoring strategy training similar to this study.

Strengths and Limitations of Nelson and Narens's Theory

In their influential work which is now considered "a classic theory of metacognition" (van Overschelde, 2008, p. 47), Nelson and Narens (1990) warned that a person's monitoring could possibly either miss some aspects of the incoming information or add aspects that are not actually present. This issue has implications for calibration research, because "currently, students' metacognition is being assessed predominantly diagnostically" (Desoete, 2008, p. 191) in one snapshot, not over time.

Considering that calibration refers to the extent of the accuracy of monitoring judgements (Bol et al., 2010) I contend that (a) any flaws in students' monitoring processes might affect their calibration and (b) classroom interventions designed to develop monitoring might possibly develop students' calibration as well. Earlier, Bateson (1972) conceptualised control processes as the thermostat on a furnace with an on-off switch, which would come online each time that students need to intensify their efforts to perform a task. A critical difference exists, however, between Bateson's furnace metaphor and Nelson and Narens's (1990) model in which the input from the meta level not only informs the object level of the necessity of some action, but also directs the object level mechanism. To connect with the thermostat metaphor, in Nelson and Narens's model, the switch does not just turn the furnace on, it intervenes to make the furnace work effectively by using the control processes to guide study behaviour. As Hadwin and Webster (2013) elaborated, Nelson and Narens's (1990) model suggests that being able to monitor the ongoing cognitive activities is necessary for the effective control of study behaviour, including the appropriate allocation of time and attentional resources. This issue might be critical for students (Winne & Jamieson-Noel, 2002), because if they base their study behaviour on miscalibrated judgements, their school performance might suffer. For instance, students who believe they have studied and practised enough for their test might terminate their test preparation prematurely.

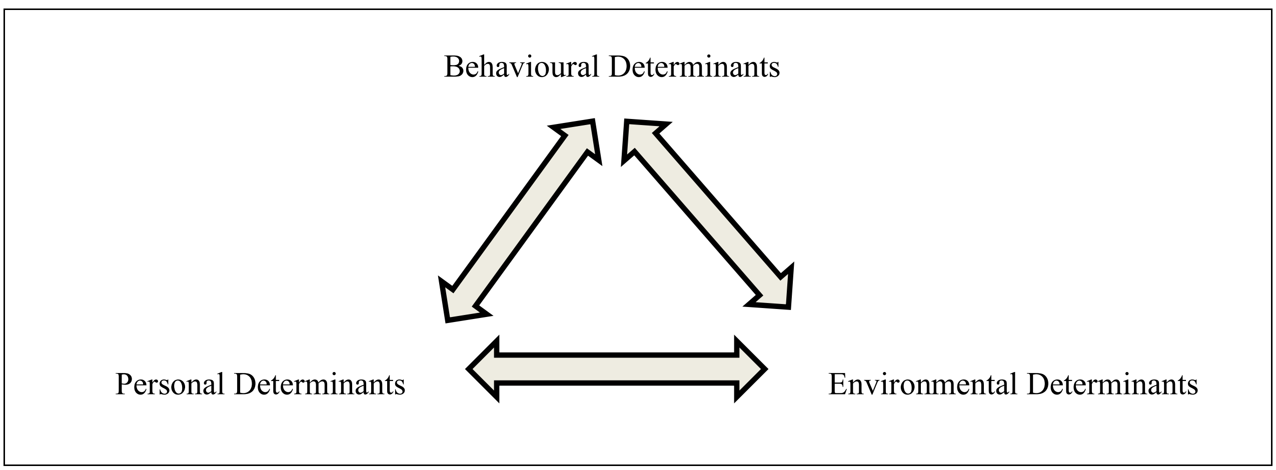
Although Nelson and Narens's (1990) theory is elegant and parsimonious, which adds to its utility, some notable limitations of the theory arise from (a) its cause and effect nature, which implies that monitoring will necessarily lead to control; (b) the assumption that the monitoring and control processes work in isolation; (c) the absence or obscurity of other metacognitive constructs such as the aforementioned planning, and evaluation; and (d) its disregard for the influence of social and affective factors. Nelson and Narens acknowledged yet another important limitation caused by individuals' erroneous monitoring judgements, as they might discard valuable information or include information which is not pertinent. Moreover, the theory neither distinguishes between the ways in which correct and incorrect judgements emerge from the meta level nor does it consider the role of personal volition and motivation. In addition to the above-mentioned limitations, Nelson and Narens proposed that despite the variability in monitoring accuracy across populations and life circumstances, considering the inevitable social and environmental influences on individuals, bias in judgement might prove unavoidable.

The Social Side of Metacognition

Considering the absence of the social aspects of metacognition in Nelson and Narens's (1990) theory, I considered it to be meaningful and appropriate to undertake this classroom study of calibration within Albert Bandura's (1986) social cognitive theory. In *Social Foundations of Thought and Action: A Social Cognitive Theory,* Bandura (1986) proposed a perspective on human development that afforded a central role to cognitive, vicarious, self-regulatory, and self-reflective processes. Critical to this view is the notion of reciprocity; Bandura theorised human functionality as a dynamic system in which personal (e.g., cognitive, affective, and biological phenomena), behavioural (e.g., personality, habits, and attitudes), and environmental (e.g., external influences, classroom atmosphere, teachers, and peers) influences interact in a triadic reciprocity. Figure 2.5 represents Bandura's triadic reciprocal determinism. Since the term 'determinants' may possibly imply causation, in placing my study within Bandura's (1986) social cognitive theory, I followed the example of Pajares (2002) and replaced the term 'determinants' with 'influences'.

Figure 2.5

Bandura's (1986) Triadic Reciprocal Determinism



The reciprocal nature of the personal, behavioural, and environmental influences posits that the efforts aimed at teaching and counselling students could be fruitful, but as noted by Pajares (2002) it poses the challenge to teachers to assist their students at the personal (e.g., building students' metacognition), behavioural (e.g., improving students' study habits), and environmental (e.g., ensuring an inclusive and supportive classroom atmosphere) levels.

Since its inception in 1986, Bandura's social cognitive theory has evolved greatly, adding many new frontiers (e.g., agency and autonomy) to the dimensions of human adaptation, change, and development. In many ways, the new and evolved social cognitive theory frames this study well. For instance, in methodology, design, and implementation, my study encouraged agency, which I understand as self-reflection and distributed control (Thomas, 2003) for students within the social setting of a classroom, without which the development of students' calibration may not be achievable. In Bandura's (2001) view, agency means to intentionally make things happen through one's actions. By referring to the social cognitive theory as an agentic theory, Bandura (2001) provided a perspective of the theory in which human agency is characterised by the core features of intentionality, forethought, self-reactiveness, and self-reflectiveness. According to Bandura (2001), "exercising control over the nature and quality of one's life is the essence of humanness" (p. 1). Bandura (2001) posited that within the small social setting of a classroom, other than personal agency, students enjoy two other kinds of agency which might support their learning behaviour and task performance. According to Bandura (2001) these two other kinds of agency are i) proxy agency in a sense of relying on others to act on their behalf to achieve results, and ii) collective agency exercised through socially coordinated efforts. In addition, regarding metacognition and metacognitive skills, such as calibration, Bandura contended that people use self-reflection to mediate between *knowing* and *doing* and called the broader notion of reflective self-consciousness as "the hallmark of human agency" and a characteristic that is "distinctly human" (p. 21), for it endows individuals with the capability to engage in reflection and self-evaluation, which may be consequential to their subsequent behaviour. Bandura (2001) posited that "the metacognitive capability to reflect upon oneself and the adequacy of one's thoughts and actions is another distinctly core human feature of agency" (p. 10). Bandura argued that the verification of the soundness of one's thinking helps one judge the correctness of one's predictive and operative thinking against the outcomes of one's actions. This important statement emphasises the critical role of metacognitive strategies, including monitoring strategies, in pursuing desired academic outcomes, theorising that by learning effective metacognitive strategies students may become more self-reflective, better engage in self-evaluation, and hence benefit from the potential modifications they may decide to make to their study behaviour.

The second connection between my thesis and the social cognitive theory rests on my contention that what students are confident about might affect their knowledge construction and motivation. Some mathematics educators might wonder why the domain knowledge of mathematics is not enough for a motivated and successful experience for mathematics students. In social cognitive theory, knowledge alone is often regarded as a poor predictor of academic task performance, because the beliefs people hold about their abilities are immensely powerful in guiding the ways in which people behave. Ferrari and colleagues (2010) posited that social cognitive theory views students as metacognitive self-reflective beings who cognise, theorise, and create as opposed to those theories in which students acquire knowledge by reacting to their surrounding environments and life circumstances. Bandura (1986) emphasised the strong influence of students' thoughts and beliefs on their aspirations, goals, and actions.

The third connection between this study and Bandura's social cognitive theory lies in the notion of self-efficacy. Since the 1990s, self-efficacy has constituted one of the most prominent aspects of the social cognitive theory. According to Bandura (1997), self-efficacy is "the beliefs in one's capabilities to organise and execute the courses of action required to produce given attainments" (p. 3). In the field of calibration research, some researchers (e.g., Hacker et al., 2008b) have argued that students' monitoring judgements stem from their self-efficacy beliefs.

In my study of calibration, because the students made their monitoring judgements before the receipt of their test performance scores, the prediction of a performance score (e.g., 85%) was an expression of the students' beliefs in their own capability to produce such a result (i.e., a self-efficacy belief). Hence, these judgements fit within the social cognitive theory, because one way or the other, these beliefs influence the course of action students choose to pursue. Bandura (1986) suggested that the "reasonably accurate appraisal of one's own capabilities is, therefore, of considerable value in successful functioning" (p. 393). Most classroom interventions, including mine, ultimately promote adaptive change in students' behaviour such as helping students identify efficient study habits, hence making social cognitive theory a viable theoretical framework. In sum, Nelson and Narens's (1990) theory of metacognition and Bandura's (1986) social cognitive theory played different but complementary roles in theoretically framing my study. While the former helped me explore the mechanics of *how* the students' calibration changed and what drove the change, the latter was useful to me in theorising *why* students' calibration changed and what influenced it.

Epistemological and Ontological Orientation of This Study

In my engagement with this intervention study of calibration, I adopted a social constructivist view as the epistemological perspective that might inform and guide my research. The constructivist perspective, partly derived from the seminal work of Jean Piaget (1970), assumes that knowledge is not passively acquired but actively constructed based on one's own experiences. As an important metaphor for the conceptualisation of human learning and development (Gergen, 1985), constructivism emerged in the 1970s and as noted by Fosnot (1996), it has been ever since used both as a theory of knowledge and a theory of learning.

Despite being central to some of the reform efforts in mathematics education (e.g., NCTM, 2000; 2007), constructivism is a term that has many shades and layers of meaning.

As argued by Schunk (2020), Bandura's (1986) social cognitive theory is compatible with social constructivism in many ways. For instance, as a theory of knowledge, social constructivism acknowledges the importance of self-regulation, emphasising individuals' capability to guide their own thoughts, emotions, motivations, and actions. Karpov (2014) provides the example of Vygotsky's Zone of Proximal Development in which interaction with others, causes individuals to evaluate the adequacy of their own knowledge and understanding. Hence, relying on their self-regulation, they may seek out the assistance of more knowledgeable individuals to improve their own knowledge and understanding. Self-regulation is a principal component of Bandura's (1986) social cognitive theory, reflecting individuals' ability to organise and conduct specific courses of action to attain their goals. A second parallel between the social constructivist thought and social cognitive theory rests on envisioning individuals as agents of their own learning. Bandura (1986) contended that modelling and imitation alone cannot sufficiently explain learning. Hence, he increasingly focused on cognitive constructs and explained the role of agency as individuals draw upon their own varied cognitive processes, including memory, language, evaluation, and anticipation, to integrate and internalise their experiences. As discussed by Schunk (2020), this perspective on agency as determined by social and biological elements and yet, emergent and not limited to these elements, is shared by both social constructivism and social cognitive theory. The notion of environment constitutes the third parallel between social constructivism and social cognitive theory. In social cognitive theory, learning does not solely depend on the environment. As I noted earlier in this chapter, the theory explains human behaviour in terms of the triadic reciprocal influences of personal, behavioural, and environmental factors. Similarly, as elaborated by Schunk (2020), social constructivist theories propose that individuals are possibly aware when they cannot reconcile their current state of knowledge and understanding with the new information presented to them by their environment and should be allowed to choose a suitable environment that supports their learning. In my study, my *pedagogy* reconciled my personal epistemology of social constructivism with my theoretical framework of social cognitive theory. To explain, in my classroom practices I encouraged student agency, self-regulation, self-reflection, social interactions, and a supportive learning environment which are promoted by both of these two different but compatible theories. Further, the terminology used in describing these two theories are compatible, as well. By defying the behaviourist terms such as responding, shaping, and chaining, Bandura (1986) utilised terms such as building, creating, reflecting, and interacting which promote agency and active construction of knowledge central to social constructivist thought.

Throughout this study, in my attempts to help students become well-calibrated self-evaluators, the ontological orientation of my work was multi-layered as it reflected my own sense of being and my personal experience with mathematics—not as an object in front of me or against me, and not as a school subject that I *had* to study, but rather as a topic worthy of discussion and profound exploration, as a way of experiencing and understanding phenomena, and as a way of describing the world and being in the world. I acknowledge that within the small social community of a mathematics classroom, as long as I am *the teacher*, my ontological orientation—who I am, what constitutes my reality, and how I see mathematics—continues to shape my practice and influence the students.

I have often wondered whether I have chosen mathematics, or it has chosen me. Palmer (2007) insightfully delineated this issue for me by stating that "as teachers, we are drawn to a body of knowledge because it sheds a certain light on our identity and the world. We do not merely find a subject to teach—the subject also finds us" (p. 26). In me, mathematics provokes a sense of self that was lying dormant in me before I discovered the ways in which it named and framed the world. As a child, my family moved around a lot, and each time that we moved, the landscape changed, the people around me changed, and so did the language they spoke and the customs and traditions they followed. When I went to school in my new city, mathematics was the familiar part of my life that never changed. Growing up in an ever-changing world that seemed uncertain, I thrived in the strength and stability of mathematics. It became my refuge and all that I needed it to be: strong, stable, welcoming, truthful, present, clear, original, coherent, creative, objective, meaningful, lively; Alive!

I concur with Palmer's (2007) view that we teach who we are. In teaching mathematics, I also teach who I am. While sharing the experience of being in a classroom with my students, as our realities merge and manifest together, I am grateful for the opportunity to introduce mathematics as a lively discipline worthy of attention and deep examination and for any instances that I can bring mathematics into my students' lives for yet another inspiring conversation. As such, my ontology is one of joy and gratitude for the privilege to encourage mathematics students to open up to new horizons, to imagine, to experience, to take a chance, to push the boundaries, to create, and—above all—to find in them a place of joy, confidence, and strength that binds them to the world and gives them the courage to stand up and *question*.

Perspectives on Metacognition From Mathematics Education and Other Disciplines

The wealth of research on metacognition in education speaks to more than four decades of work and many attempts (e.g., Desoete & Veenman, 2006; Palincsar, 1986; Palincsar & Brown, 1984; Paris & Oka, 1989; Pressley & Afflerbach, 1995; Roderer & Roebers, 2013;

Thomas, 2003, 2006; Thomas & Anderson, 2014; Thomas & McRobbie, 2001) to apply the metacognition theory to educational settings.

Educational psychology enjoys a long-lasting tradition in metacognition research, with many contributing scholars (e.g., Alexander, 2013; Bol & Hacker, 2012; Brown, 1980; Chen, 2002; Desoete, 2008; Dunlosky & Thiede, 2013; Efklides, 2011; Koriat, 2011; Kruger & Dunning, 1999; Lin & Zabrucky, 1998; Metcalfe & Finn, 2008; Nelson & Narens, 1990, 1994; Nietfeld & Schraw, 2002; Pajares, 2002; Palincsar, 1986; Paris & Oka, 1989; Parkinson et al, 2010; Pintrich, 2000; Schneider & Artelt, 2010; Schraw, 1995; Schunk, 1984, 2008; Stone, 2000; Tobias & Everson, 2000, 2002; Veenman & Spaans, 2005; Winne, 2001, 2004; Zabrucky, 2010; Zimmerman, 1983) who have attempted variously and to varying extents, to bridge the educational theory and practice by applying the metacognition theory to classroom settings. Palincsar and Brown's (1984) *reciprocal teaching*, a procedure used to enhance students' strategic reading skills, and Zimmerman's (2000) theory of academic *self-regulation*, which involves three sequential phases of forethought, performance, and self-reflection are among these efforts.

Metacognition in Mathematics Education

In mathematics education, some researchers (e.g., Goos, 1994; Schoenfeld, 1987b; Stacey, 1991) have attributed students' success in different areas of mathematics to students' metacognitive strategies. Particularly, it has been widely noted (e.g., Desoete et al., 2003; Dignath & Buttner, 2008; Goos, 2002; Montague, 1998; Schoenfeld, 1992, 2007; van der Stel & Veenman, 2010; Veenman et al, 2005) that metacognition influences all the aspects of mathematical problem-solving at all levels of difficulty. In my experience, metacognition monitoring skills might potentially prevent blind calculation, false generalisation, or superficial number crunching approaches such as answering 24 to the question '20 is four more than…' because *more* means addition.

Garofalo and Lester (1985) asserted that despite the importance of metacognitive processes in mathematical performance, metacognition is not systematically studied within the mathematics education community for at least three reasons:

> First, covert behaviour of any type is extremely difficult to observe and analyse. In fact, many psychologists believe that people have no direct access to their mental processes, thereby making self-reports highly suspect as a source of data. This suspicion is especially pronounced in the case of metacognitive processes. Second, researchers who accept self-reports as legitimate data recognize that asking a person to verbalize information while performing a task may affect cognitive processes if the verbalized information would not otherwise be attended to. Third, phenomena linked with metacognition, metamemory, and the like have been regarded by many psychologists as being too ill-defined to be the objects of scientific investigation. (pp. 166–167)

Garofalo and Lester (1985) further suggested that "since mathematics education researchers have often followed the lead of experimental psychologists in deciding what to study and how, the role of metacognition in mathematical performance has been overlooked" (pp. 166–167).

Although the above comments were made more than 3 decades ago, the critical role of metacognition and its components in mathematical task performance has not yet been sufficiently and systematically explored in mathematics education research, which might possibly explain why the concept of calibration, to my knowledge, has not appeared anywhere in mathematics education research literature. In my view, calibration, which is a measure of the accuracy of students' monitoring judgments, enacts the role of metacognition in mathematical task performance.

The construct of metacognition itself has been attended to by many mathematics education researchers, including Biryucov (2004), Cai and Lester (2010), Campione et al. (1989), Cardelle-Elawar (1992; 1995), Lesh (1982), Lester and Garofalo (1982), Schoenfeld (1981; 1982; 1987b; 1998; 2010; 2011), and Silver (1982a; 1982b). In fact, from the early 1980s, researchers with a focus on mathematical problem-solving demonstrated an interest in metacognition. For instance, Lester (1982) maintained that a person's metacognitive knowledge and monitoring capability can substantially affect successful problem-solving in mathematics.

Silver (1982b), a strong proponent of the role of metacognition in mathematics education, contended that successful problem-solving may solely depend on students' metacognitive skills. While Verscaffel (1999) considered metacognition to be of particular importance in mathematical problem-solving, Garofalo and Lester (1985) argued for the essential role of metacognition in the understanding and analysis of mathematical task performance. Interestingly, without using the term metacognition, Tall (2013) posited a monitoring approach in dealing with what he calls "problematic met-befores," suggesting that to "rationalize a new problematic situation, it may be helpful to encourage the learner to recall a situation in which the idea worked and continue to work" (p. 88), which constitutes a well-known and well-documented (Schoenfeld, 1979, p. 178) monitoring strategy.

Alan Schoenfeld (e.g., 1979, 1981, 1982, 1985, 1987a, 1987b, 1992, 1998, 1999, 2000, 2002, 2005, 2006, 2009a, 2010) is among the most prolific and influential scholars who have attempted to bridge metacognition research and classroom teaching of mathematics for almost 30 years. In response to several mathematicians who questioned the excessive interest in metacognition, Schoenfeld (1987b) wrote a chapter titled *What is all the fuss about metacognition?* in which he challenged the then-prevalent view of teaching mathematics as a set of definitions, theorems, formulas, and algorithms, and made a compelling argument for using metacognitive strategies in teaching mathematics for a meaningful learning experience. Many mathematics education scholars, including Silver (1982b), Pressley (1986), and Schoenfeld (1992), have grounded their approaches to teaching students metacognitive strategies in the foundational work of Polya (1957, 1973). Although Polya did not introduce his problem-solving heuristics as metacognitive strategies, he highlighted the importance of self-monitoring and thinking about one's own thinking in his influential work.

Intervention Studies in Mathematics Classroom Settings

In this section, I review the intervention studies conducted in mathematics classroom settings, which—like my study—have focused on the classroom applications of metacognition. From the early 1980s, some researchers in the area of mathematical problem-solving (e.g., Lester, 1982, Silver, 1982a, 1982b) became interested in the applications of metacognition in a classroom setting. For instance, Garofalo and Lester (1985) revised Polya's influential four-phase problem-solving heuristic to properly incorporate metacognitive strategies. Garofalo and Lester (1985) proposed a cognitive-metacognitive theoretical framework for learning mathematics, which involved an orientation phase (i.e., the strategic behaviour to actively assess and understand a problem), an organisation phase (i.e., the planning of a course of action), an execution phase (i.e., the regulation of behaviour to carry out the plan), and a verification phase (i.e., the evaluation of decisions and outcomes), each enriched with both cognitive and metacognitive activities.

With a focus on mathematical thinking and problem-solving for university students, Schoenfeld (1987b) insisted that as a living, evolving discipline, mathematics has the potential to offer students lasting and meaningful learning experiences and that metacognition is best fostered within a mathematics culture where students engage in communicating, practicing, and building their mathematical knowledge and thinking.

For the purpose of his pedagogical intervention, Pressley (1986) devised the Good Strategy User Model in which the term 'strategy' has a broad meaning similar to procedural metacognitive knowledge. Based on Good Strategy User Model, Pressley introduced the following five main principles of mathematics instructions to be used in a classroom context: (a) the specific teaching of mathematical strategies, self-testing, and monitoring, (b) teaching students when, where, how, and why to use specific strategies, (c) teaching strategy knowledge so that students use the right strategy, (d) the enrichment of students' knowledge base (e.g., the knowledge of the multiplication tables) towards the ultimate automatic performance of basic mathematical operations, and (e) allowing students to practice each concept separately and enough times before combining it with other concepts.

In their classroom intervention, Allardice and Ginsberg (1983) explored the development of mathematical thinking, concluding that students' performance might suffer in the absence of the control processes to guide their thinking and understanding. This view leads to the question: Is metacognitive knowledge all that students need to be successful in their respective fields? Many researchers (e.g., Kramarski & Mevarech, 2003; Kuhn, 2016; Schraw, 1998) have demonstrated that in addition to metacognitive knowledge, teachers should explicitly teach students monitoring strategies. Further, for mathematics classroom interventions to be effective, Kuhn (2016) recommended that in order to develop students' metacognition, the instructions for using metacognition should be delivered at the meta level (i.e., they should be about planning, monitoring, control, and evaluation) rather than at the students' performance level. For instance, students should be encouraged to plan a solution and monitor every step of their work while solving an equation or seek out ways to double-check their final answers.

The question of how to improve mathematics students' metacognitive skills was explored by Carr and colleagues in 1994. The researchers found that the use of internal mental strategies, such as counting in one's head, but not external strategies, such as counting on one's fingers, was related to improved metacognitive skills. In 1995, Carr and Jessup successfully replicated the positive results gained by Carr et al. (1994), demonstrating that students' metacognitive knowledge could substantially influence the development of students' metacognitive strategies.

In an experimental intervention in 1995, Cornoldi and colleagues focused on developing mathematics students' metacognitive knowledge and control processes and found that not only in normally achieving children enhanced metacognition was related to academic achievement in mathematics, but also children with a learning disability benefitted considerably from metacognition training.

More recently in Germany, Cohors-Fresenborg and Kaune (2001) and Kaune (2006) took the novel approach of designing a mathematics curriculum for school-age children using a system that used age-appropriate metacognitive activities. These activities focused on enhancing students' self-reflection processes by encouraging a discourse-based teaching culture. As such, through relevant classroom activities and conversations, metacognitive strategies were embedded in students' daily learning routines. Researchers maintained that the value of this approach in improving both students' metacognition and their understanding of mathematical concepts had been demonstrated.

In 2001, while conducting research with elementary school children, Desoete and colleagues reported that students' individual differences in terms of their metacognitive prediction and evaluation skills differ among students according to their varying levels of achievement. Following this line of research, Desoete et al. (2003) conducted an intervention study with third graders which involved prediction and evaluation tasks on mathematical problem-solving both before and after task performance. The researchers demonstrated that the students who received metacognitive strategy training showed significant gains in metacognitive skills as well as problem-solving task performance.

In another classroom intervention involving a metacognitive strategy training program called IMPROVE (an acronym for Introducing new material, Metacognitive questioning, Practising, Reviewing, Obtaining mastery on cognitive processes, Verification, and Enrichment and remedial), Kramarski and Mevarech (2003) provided eighth grade mathematics students with a set of metacognitive questions. These questions included comprehension questions (e.g., What is the problem all about?), strategy questions (e.g., What kind of strategies can you use to solve this problem and why?), connection questions (e.g., How is this problem related to other problems you have already solved?), and reflection questions (e.g., Does the solution and the final answer make sense?). The comprehension questions aimed to encourage students to think about what strategies might be appropriate for a certain task and the reason why they thought so. The strategy questions inquired about the appropriateness of particular strategies, and the connection questions were meant to encourage students to identify the deep mathematical structure of a task so they could call upon relevant prior knowledge and strategies. The results revealed that IMPROVE positively affected students' metacognition including their planning and comprehension processes. Emphasising the social aspects of metacognition, Kramarski and

Mevarech (2003), who used collaborative groups in their study asserted that the social setting helped students working in groups express their mathematical thinking more ably than those who were working alone.

In an intervention study with school-age children in the Netherlands, van Luit and Kroesbergen (2006) focused on self-instruction in mathematical problem-solving, a method they designed for children with learning disabilities in mathematics. During the 16-week-intervention, the children in the training group received several training sessions on how to control their learning activities (e.g., check the strategy in use and check the answer). The researchers demonstrated that the children in the training group performed considerably better than the control group, and the finding remained stable during the post-intervention follow-up period.

Schneider and Artelt (2010) point out that although the research on metacognition in mathematics education continued after the 1990s and in the new millennium, most of the research done with children and adolescents focused only on the predictive potential of metacognitive knowledge but not on monitoring or control processes. Schneider and Artelt provided an example of their own longitudinal research project, highlighting the interrelations of metacognitive knowledge and prior knowledge in different subject domains, including mathematics. The first assessment was done in Germany with 9- and 10-year-olds in 2008, in mathematics. Schneider and Artelt (2008), who have predominantly focused on the developmental aspects of metacognition, assert that the older students in the group knew more about metacognitive strategies and performed better on mathematical tasks than their younger peers. Further, Schneider and Artelt (2008) argued that considering only the data from cross-sectional studies might not be sufficient for researchers to exclude the possibility of unknown factors affecting students' metacognition. Hence, Schneider and Artelt (2008) proposed that intervention studies that take place over many sessions might be more adequate to explore the influence of any strategy training on students' metacognition.

Although experimental studies in metacognition research are necessary and insightful (Nelson & Narens, 1994), due to the complexities inherent in conducting research on metacognition with students in the natural setting of school classrooms, it is conceivable that purely experimental research designs might not quite adequately capture the dynamic nature of metacognitive processes, including calibration. However, despite the complexities, my review of the intervention studies in mathematics classrooms revealed that well-designed and well-implemented classroom interventions have the potential to improve mathematics students' metacognition.

Calibration: Accuracy of Monitoring Judgments

As noted earlier, metacognition broadly refers to the knowledge about, and monitoring and control of cognition (Paris & Oka, 1989; Pintrich et al., 2000; Schneider & Artelt, 2010; Winne & Hadwin, 1998). Tobias and Everson (1996, 2002) contend that among the different components of metacognition, monitoring processes are basic to the others, because unless students are able to monitor their thinking activities and make relatively accurate self-evaluations of their task performance, they may not decide to either regulate or control those activities. To clarify, monitoring processes, which inform the control processes (Nelson & Narens, 1990), are described as one's awareness of task performance while on a task (Schraw & Moshman, 1995). For instance, while engaged in solving limit problems in calculus, unless students recognise that they have not grasped the concept of limits of functions, they might not spend much time and effort trying to learn and practise this concept.

Different categories of monitoring judgments were discussed earlier (see Table 2.1). As noted by Hadwin and Webster (2013), these judgments are indicative of an individual's monitoring processes. In my study, students' retrospective monitoring judgments, also called confidence judgments of performance (Schraw 2009), after completing a task, were explored. For instance, after writing a test, a student might think, "I am confident that I will get 80%," which is called a retrospective confidence judgment (Nietfeld & Schraw, 2002), indicating the level of the student's confidence in task performance. If this student scores 80% on the test, he or she is said to be well calibrated. A score lower than 80% indicates student's positive bias or overestimation (Chen, 2002; Mengelkamp & Bannert, 2010), whereas a score higher than 80% is indicative of negative bias or underestimation.

Calibration research began with experimental studies in psychology in the 1950s, but researchers' interest in the topic can be traced back to Woodworth and Thorndike's (1900) paper on judgments of magnitude, half a century earlier. As explained in a review by Kahneman and colleagues (1982), in the 1950s, the concern was that medical doctors were consistently overconfident in judging the accuracy of their diagnoses. Researchers then contrasted the calibration of medical doctors with that of other professionals and soon, as highlighted by Lichtenstein et al. (1982), turned their attention to adults' calibration on general knowledge tasks.

Like metacognition, despite the lack of a consensual conceptualisation, the construct of calibration has been studied by cognitive, developmental, experimental, and educational psychologists for decades (e.g., Alexander, 2013; DiFrancesca et al., 2015; Fin & Tauber, 2015; Fischhoff et al., 1977; Glenberg & Epstein, 1985; Gutierrez et al., 2016; Koriat, 2011; Muis et al., 2016; Nietfeld et al., 2006b; Pajares & Kranzler, 1995; Reid et al., 2017; Schraw et al., 1993), within and beyond a variety of academic domains such as psychology (Hacker et al., 2000; Pintrich, 2004), human development (Dinsmore & Parkinson, 2013), business studies (Stoten, 2019), medical sciences (Persky & Dinsmore, 2019), general literacy skills (Maki et al., 2005), reading comprehension (Schraw et al., 1993), decision-making (Callender et al., 2016), reading at elementary school level (Huff & Nietfeld, 2009), and mathematics (Barnett & Hixon, 1997; Garcia et al., 2016; Kramarski & Mevarech, 2003; Rutherford, 2017a).

Educational Importance of Calibration

Flavell (1979) stated, "I find it hard to believe that children who do more cognitive monitoring would not learn better both in and out of school than children who do less" (p. 910). Researchers (e.g., Alexander, 2013; Bol & Hacker, 2001; Flavell, 1979; Kruger & Dunning, 1999; Nietfeld et al., 2005) have long demonstrated that for a variety of reasons, people in general, including students, are quite limited in their knowledge and monitoring of their own cognition. In other words, they are not well-calibrated. In this regard, Nelson and Narens (1994) asserted, "People are construed as imperfect measuring devices of their own internal processes" (p. 18).

Koriat and colleagues (2002) pointed out that there often exists a difference between students' subjective self-evaluative feelings on a topic and their objective performance outcomes on the same topic. This discrepancy can give rise to the illusion of knowing (Glenberg & Epstein, 1985; Kruger & Dunning, 1999), in which students assume they are prepared for a test, but they are not. As a mathematics teacher, I am familiar with students' illusion of knowing and its negative consequences for students' motivation and achievement. I think it is important for mathematics teachers to try to understand the basis of students' monitoring judgments, what affects the accuracy of these judgments, how these judgments can be matched with students' objective task performance outcomes, and how students' calibration can be developed in the context of day-to-day classroom activities.

A Problem of Conceptualisation

Within the calibration research literature the lack of a clear definition has been problematic because as noted by Schunk (2008), how a concept is conceptualised affects how it is investigated and assessed and how the research results are understood and interpreted. In educational psychology, one definition of calibration (Schraw et al., 1993) is the accuracy of judgments in evaluating task performance. Some researchers, however, have suggested much broader definitions for calibration. For instance, Glenberg et al. (1987) and Wagenaar (1988) defined calibration as a measure of the relationship between confidence in performance and the accuracy of performance, while Lichtenstein et al. (1982) proposed that calibration is the degree to which individuals' judgements about their understanding, capability, competence, and preparedness corresponds to the understanding, capability, competence, and preparedness they actually manifest. To provide a more recent example of how calibration is conceptualised, Crane and colleagues (2017) define calibration as a skill of monitoring, referring to students' ability to monitor their understanding and task performance.

Within the research literature, calibration has been conceptualised predominantly numerically as a correlation (Fischhoff et al., 1977), a relationship (Stoten, 2019), a measure (Bjorkman, 1992; Cooke, 1991; Glenberg et al., 1987; Sharp et al., 1998; Wagenaar, 1988; Yates, 1990), a degree (Keren, 1991; Lin & Zabrucky, 1998), a comparison (Gutierrez, 2017), the accuracy of judgments (Rutherford, 2017a), an absolute value (Maki et al., 2005), the difference (Bol & Hacker, 2012), and the extent of congruence (Gravalia & Gredler, 2002) for the past 4 decades, which partly explains why researchers have predominantly used quantitative and experimental research methods to investigate calibration and why the personal, social, and affective factors are by and large absent from calibration research. Further, calibration provides a good example for Schunk's (2008) argument in which inconsistencies in conceptualisation have led to inconsistencies in research findings.

I contend that the ways in which calibration is conceptualised, and how these conceptualisations have changed and evolved are important to my understanding and my review of research literature on calibration. As such, I have summarised the evolution of the conceptualisation of calibration in research literature over the past four decades in Appendix A.

Approaching a Working Definition for Calibration

In my search for a working definition of calibration which might best suit my research, I considered the conceptualisation suggested by Lichtenstein et al. (1982), which describes calibration as the degree to which individuals' judgments about their understanding, capability, competence, and preparedness correspond to the understanding, capability, competence, and preparedness that they actually manifest. Calibration researchers have often perceived students' performance judgments as a reflection of their subject-matter understanding (Gutierrez & Schraw, 2015; Huff & Nietfeld, 2009, Lindsey & Nagel, 2015), capability (Sheldrake et al., 2014; Stoten, 2019), competence (Persky & Dinsmore, 2019; Sheldrake et al., 2014), and preparedness (Ots, 2013; Thiede & Leboe, 2009). Due to the interpretive nature of my research and my choice to employ both qualitative and quantitative methods, I opted for 'the extent of agreement' instead of 'the degree of correspondence,' and thus operationalised calibration as *the extent to which individuals' judgments about their understanding, capability, competence, and preparedness agree with the understanding, capability, competence, and preparedness they actually manifest.* This definition of calibration takes into account the personal influences such as students' beliefs, the behavioural influences such as their attitude toward test preparation, and the environmental influences such as the availability of a metacognitively supportive learning environment on students' calibration. As such, this definition of calibration is consistent with Bandura's (1986) triadic reciprocal influences of personal, behavioural, and environmental factors.

Classroom Interventions Aimed at Developing Students' Calibration

The fundamental paradigm shifts in educational research during the late 1970s and the early 1980s changed the focus of the discipline from condition to cognition (Zimmerman, 2002), and presented a new perspective on students' individual differences through the research on social cognition and metacognition. In this line of work, intrigued by students' learning limitations related to the lack of metacognitive awareness and inability to compensate for those limitations, researchers (e.g., Schunk, 1989; Zimmerman, 1989) attempted to raise students' metacognitive awareness of the subject-matter content by encouraging them to set goals and self-monitor the steps taken toward those goals. Prior to such research, Shapiro (1984) had demonstrated that simply by asking students to set a goal and monitor and self-record some aspects of their learning, an intervention in and by itself, students showed superior academic gains. This educationally interesting phenomenon, called 'reactivity' (Shapiro, 1984; Zimmerman, 2001, 2002), suggests that developing students' monitoring skills might support their academic gain. Zimmerman theorised that although the lack of fundamental skills may pose a problem, reactivity can potentially produce a readiness that is essential for personal change, supporting the effectiveness of almost any kind of pedagogical intervention.

Students' expectations and performance judgements have been linked to better study habits (Horgan, 1990) as well as to improved task performance (Bol et al., 2005; Butler &

Winne, 1995; Chen, 2002; Schraw et al., 1993) in academic domains. Within the calibration research literature, most studies have focused on improving students' academic task performance (e.g., Bol et al., 2005; Bol et al., 2012; Nietfeld et al., 2005, 2006a; Rutherford, 2017a, 2017b; Tobias & Everson, 1996, 2000, 2002; Zimmerman & Moylan, 2009) and are not strictly centred on improving students' calibration accuracy. Moreover, as noted by Bol et al. (2012), calibration studies have predominately defined calibration as a stable psychological construct. This view may be partly responsible for the dominance of snapshot-style experimental research, which has likely led to some inconclusive or contradictory results. Consistent with Bandura's (1986) social cognitive theory, a different insight provided by some researchers (e.g., Bol et al., 2005; Dembo & Jakubowski, 2003; Hacker et al., 2000; Thomas, 2002, 2006; Thomas & McRobbie, 2012; Roeser et al., 1998) suggests that metacognition might not be purely psychological but rather a socially constructed and culturally mediated construct. Calling on contemporary researchers to abandon examining metacognition solely as a reflection of one's cognitive ability, Zimmerman (1990) convincingly argues that the research on metacognition must also include the complex social dimensions of motivation, emotion, and behaviour, as these factors account for the necessary humanistic nature of metacognition. This insight is consistent with the scholarly calls (e.g., Roderer & Roebers, 2013) for conducting calibration research in naturalistic classroom contexts.

A line of research pursued by Bol, Hacker, and their colleagues was carried on between 2000 and 2012. Hacker et al. (2000) started out by asking an important question: Can students accurately predict their task performance outcomes? The researchers also asked the following questions: Does accuracy vary with performance, and does prediction accuracy increase over multiple tests? Ninety-nine students in an undergraduate educational psychology course participated in a semester-long classroom intervention, receiving information on the importance of self-monitoring and accurate self-assessment throughout the course. They also took a practice test before each of their three tests and were encouraged to focus on their mistakes. Students received performance feedback on their grades after each test. Hacker et al. (2000), who analysed students' predictions and postdictions by utilising calibration graphs, concluded that high-achieving students were highly accurate and their calibration accuracy improved, while low-achieving students were strongly overconfident, overestimating their performance scores, with their overestimation being greatly exaggerated the lower they performed. But why would low-achieving students repeatedly make unrealistic performance judgements despite feedback? The researchers theorised that those students believed they would do well on the test and attributed their poor test performance to external factors such as a tricky test or an unreasonable teacher. This study was unique in the sense that the participants were told to base their confidence judgements on their performance outcomes of the practice test, which might be a useful strategy, considering the variability of the factors that may potentially influence students' confidence judgements.

In 2001, by adding the element of *review* to their previous work, Bol and Hacker conducted a classroom intervention with 59 undergraduate students in a research methods course, comparing the effects of practice tests and traditional review on performance and calibration. In total, two tests and two practice tests were taken by the students. In line with previous research, Bol and Hacker determined that (a) the participating undergraduate students were not very good at predicting or postdicting their performance outcomes, (b) higher-achieving students were more accurate than their lower-achieving peers but underconfident, and (c) low-achieving students were less accurate than their higher-achieving peers in their predictions or postdictions of performance outcomes and were overconfident. Bol and Hacker concluded that giving practice tests were less useful than the traditional review by students, as the students in the study seemed to focus too narrowly on the material included in the practice tests.

In 2005, Bol and colleagues conducted another study with 356 undergraduate students to investigate the influence of overt practice and performance level on calibration accuracy and academic task performance. This time, recognising that the previous attempts to improve students' calibration accuracy had produced mixed results, the researchers included students' explanatory styles. Explanatory styles—people's ways of interpreting daily events and consequences (Ormrod, 1999)—have been linked to a wide range of academic task performance outcomes. Bol et al. (2005) wondered whether students' explanatory styles were responsible for their unrealistically optimistic judgements, even in the face of repeated failure. This study confirmed that high-achieving students were more accurate than their low-achieving peers but underconfident, while the low-achieving students, who were less accurate, were noticeably overconfident. Bol et al. demonstrated that approximately 32% of performance scores and prediction/postdiction accuracy were related to students' explanatory style, and suggested that "because calibration accuracy may be influenced by a more global self-concept of ability rather than the actual performance data, this may be one reason why calibration is resistant to change" (p. 289). The possibility that students' self-concepts of ability may influence their calibration has been taken up by other researchers, including Dembo and Jakubowski (2003), Hacker et al. (2000), and Roeser et al. (1998).

The notion that students' explanatory styles may affect their calibration accuracy prompted a study by Hacker and colleagues (2008a), in which they used a two-by-two quasi-experimental design to investigate the impact of extrinsic incentives and reflection on students' calibration. The researchers examined the relationship between explanatory style, performance, and monitoring judgements for 137 undergraduate students enrolled in an educational psychology course. The results showed that high achievers were very accurate in their judgements. The calibration accuracy of the high-achieving group, however, did not improve substantially over the course of the semester, and their explanatory styles did not strongly affect their accuracy. Low achieving students were minimally accurate in their judgements, but those low-achieving students who received incentives showed substantial improvement in their calibration accuracy. The researchers used open-ended questions to explore the accuracy of monitoring judgements and noted that explanatory style was a powerful predictor of calibration accuracy.

Bol et al. (2010) investigated the calibration accuracy of 77 middle school students on mathematical tasks. The results were consistent with some of the previous research results with undergraduate students; high-achieving students demonstrated better calibration accuracy than their low-achieving peers. The only difference was that in this study, all students were found to be overconfident. The researchers concluded that interventions at earlier grades, such as the work done by Brookhart et al. (2004) with third graders, may play an important role in positively enhancing students' calibration accuracy. It was also demonstrated that all students attributed their monitoring accuracy to their perceived academic capabilities, as well as to their study time, personal effort, and expectations of test difficulty.

Finally, Bol and Hacker (2012) conducted another two-by-two quasi-experiment to investigate the effects of studying individually or in groups and with or without guidelines on the calibration accuracy of 82 high school biology students. The researchers demonstrated that students' calibration accuracy and performance scores improved with guidelines that helped students monitor their performance in a group setting. Students' explanatory styles were not considered in this research project. Bol and colleagues who have conducted research on calibration accuracy for more than a decade, confirm that calibration accuracy is malleable (Bol & Hacker, 2012) and subject to change under careful training—an optimistic outlook that I also embraced in deciding to conduct this study.

In a similar line of research, Nietfeld and colleagues (2005) explored whether practising monitoring strategies consistently over a semester will lead to significant improvements in students' calibration accuracy. A group of 27 undergraduate students practised monitoring strategies. The researchers examined the occurring changes in students' monitoring accuracy and its relationship with performance. The results suggested that calibration accuracy did not improve and remained stable across several tests. Contrary to Hacker and colleagues' (2000) findings, the researchers did not observe any improvements in calibration accuracy of high-achieving students, possibly due to the different approaches used to measure the change in students' calibration accuracy. While Nietfeld et al. (2005) compared the global and local judgments, Hacker et al. (2000) used pre- and post-performance judgments to conduct research. Nietfeld et al. (2005) demonstrated that in their study, the global judgements over the whole test were more accurate than the local judgements over one item, and the students' performance was strongly linked to the accuracy of their judgements. In other words, the low-achieving students were also weak at knowing where to direct their efforts to improve performance. This study was one of the first studies to explore calibration accuracy in a naturalistic classroom context over an extended period of time.

In 2006, Nietfeld and colleagues conducted a follow-up study investigating the effects of distributed monitoring exercises and feedback on the calibration accuracy, performance, and self-efficacy of 84 undergraduate educational psychology students over 16 weeks. The researchers reported significant improvement in both calibration accuracy and performance scores and concluded that even modest interventions aimed at improving monitoring may enhance students' calibration. By addressing the relationship between students' monitoring skills and motivation, the researchers suggested that it is worth examining how students' self-generated feedback affects their study strategies during the learning process.

More recently, in an experimental study, Gutierrez and Schraw (2015) explored the effects of strategy training and incentives on students' calibration. The participants who were 107 undergraduate psychology students received three sessions of instruction training in seven general areas. These areas were: reviewing the text for the main objectives, summarising, re-reading text and reflecting, using contextual cues, highlighting text, relating similar questions, and using diagrams. The researchers reported the significant effects of instruction training and incentive on calibration accuracy of the participating students.

Rutherford (2017b) analysed elementary school students' calibration through studying their interactions with a year-long digital mathematics curriculum. The findings suggested that the differences in students' calibration accuracy were related to the differences in students' academic achievement in mathematics. Further, Stoten (2019) employed *learning diaries* to qualitatively explore calibration awareness of undergraduate business students. The researcher inferred that although students might have some insights into how to approach studying a certain topic, they had a far less developed sense of how to enhance their metacognitive skills including calibration.

Despite the call for conducting research on calibration accuracy in a classroom context, across subjects and age levels, using relevant assessment material, by classroom teachers and what has been learned from those students, few intervention studies have been conducted in K-12 classrooms, and to my knowledge, none in senior mathematics classrooms (Grades 11 and 12) with the explicit goal of developing students' calibration. The intervention studies that I reviewed in my thesis present an optimistic view of calibration in which calibration accuracy is a malleable construct (e.g., Gutierrez, 2017; Gutierrez & Schraw, 2015; Stoten, 2019) and its importance lies in supporting students to become self-regulated life-long learners capable of identifying their own learning strengths and weaknesses (Stoten, 2019).

Overall, the calibration research findings suggest that calibration is a dynamic construct and far from being purely psychological. As asserted by Hacker et al. (2008b), despite decades of research on different aspects of calibration, the question of *how calibration accuracy can be improved in a classroom context* has not been definitively answered. This important and complex question that is taken up in this study has educational implications for students' learning, motivation, and achievement. Hence, identifying the contributing factors to calibration change and designing effective interventions remain fertile areas for research.

Findings, Gaps, and Contradictory Results in Calibration Research

Research Findings

With a focus on what was pertinent and informative to my study of calibration, in this section I present the research findings on calibration over the span of 48 years, between 1971 and 2019. It is important to note that these findings are based on predominantly laboratory-style studies and I could not find any studies undertaken by a classroom teacher to develop students' calibration in a high school mathematics classroom setting, similar to my study. I have mentioned some of these findings but not all in the previous sections of this document. This synthesis is non-exhaustive, as it is limited to the topics pertinent to my study.

Research Findings Linking Students' Monitoring to Their Task Performance. Research shows that accurate monitoring of task performance by students is certainly not a given (Glenberg & Epstein, 1985; Pressley et al., 1990). While low-achieving students are poorly calibrated and overconfident (Hattie, 2013; Kruger & Dunning, 1999), high-achieving students are well calibrated and underconfident (Hacker et al., 2008b). Notably, low-achieving students are not only less accurate in the estimation of their task performance than their high-achieving peers, but also less aware that they are inaccurate (Hattie, 2013; Kruger & Dunning, 1999).

Research also suggests that metacognitive abilities are malleable and independent of general intelligence (Pressley et al., 1990). Explicit monitoring is linked to improved grades at school (Johnson & White, 1971; Morgan, 1987) and a statistically significant relationship links students' postdictions of task performance to their actual task performance (Schraw et al., 1993).

Research Findings on What Might Influence Calibration. Research suggests that metacognition, metacognitive monitoring, and calibration are domain specific (Veenman & Spaans, 2005; Tobias & Everson, 1996). In the field of mathematics, researchers posit that metacognitive processes affect mathematics performance and vice versa (Legg & Locker, 2009; Vanderstel & Veenman, 2010). Further, interventions have been found to be more successful in mathematics than in other subjects (Dignath & Buttner, 2008). Puncochar and Fox (2004) suggest that social factors may potentially affect calibration. However, calibration is sometimes resistant to change. This is called the 'Problem of Resistance' (Bol & Hacker, 2001; Bol et al., 2005; Nietfeld et al., 2005). Calibration is linked to metacognitive skills and performance level (Bol et al., 2005; Rutherford, 2017a). Further, postdiction judgements made retrospectively are generally more accurate than predictions made prospectively. This phenomenon is called the 'Testing Effect' (Crissman, 2006; Schraw, 2009; van Loon et al., 2013).

Research Findings on Interventions That Aim to Improve Calibration. Calibration researchers posit that it is possible to improve calibration using an intervention (Glenberg et al., 1987; Hacker et al., 2000; Nietfeld et al., 2006a; Nietfeld et al., 2006b; Schraw et al., 1993; Yates, 1990). External incentives are assumed to improve calibration and performance, possibly only for low-achieving students (Hacker et al., 2008a). Furthermore, calibration and performance improved when calibration was practised with guidelines (Bol et al., 2012). Persky and Dinsmore (2009) propose that having students practise their calibration over time and investigating their sources of confidence can potentially produce more meaningful data on calibration accuracy. Moreover, research suggests that understanding why students adopt different and sometimes unjustified perceptions of their task performance is essential to any efforts to support students' learning and achievement (Stoten, 2019).

Research Findings on Monitoring Strategy Training. Research shows that teaching students monitoring strategies improves students' calibration (Schraw et al., 1993). However, researchers posit that monitoring instructions improve calibration and performance possibly only in high-achieving students (Hacker et al., 2000). Nietfeld et al. (2005, 2006a) propose that without explicit monitoring training students' calibration does not improve, but with explicit monitoring training students' calibration improves (Huff & Nietfeld, 2009; Nietfeld et al., 2006a). In the field of mathematics, research suggests that specific monitoring training increases calibration accuracy and directly benefits students' performance on mathematical tasks, in particular in problem-solving (Desoete, 2009; Desoete et al., 2003; Pennequin et al., 2010; Stillman & Mevarech, 2010).

Research Findings Regarding the Potential Role of Feedback and Reflection in Improving Calibration. Research suggests that providing grade feedback alone does not improve students' calibration accuracy (Schraw et al., 1993). Calibration researchers posit that continuous feedback improves students' calibration accuracy (Callender et al., 2016; Stone, 2000; Zimmerman, 1990). Flannelly (2001) proposes that feedback on prior performance outcomes improves students' calibration accuracy and Nietfeld and colleagues (2006a) suggest that both self-generated and externally provided feedback boost students' calibration accuracy. Further, Schraw and colleagues (2013) showed that student's self-generated feedback positively influenced their calibration.

Self-reflection, also called reflection for brief, is a central tenet of the models of self-regulation learning (e.g., Zimmerman, 1986) and researchers posit that reflection improves self-regulation and calibration (Schraw et al., 1993). In Bryman's (2001) study, seminar periods were set aside for university-level students for self-reflection. Moreover, Efklides (2008) suggests that metacognitive monitoring can take the form of self-reflection and support the control processes to improve calibration accuracy. Stoten (2019) recommends that learning through reflection by using diaries has the potential to improve students' calibration accuracy.

Gaps in Calibration Research

Researchers have identified several potentially fertile areas for research on calibration, including conducting research in classroom contexts (Hacker et al., 2008a; Hacker et al., 2008b; Lundeberg & Fox, 1991; McCormick, 2003; Nietfeld et al., 2005; Roderer & Roebers, 2013; Valdez, 2013; Winne, 2004), across ages and subject levels (Hacker et al., 2008b; Parkinson et al., 2010; Roderer & Roebers, 2013; Valdez, 2013), and designing effective interventions (Alexander, 2013; Barnett & Hixon, 1997; Dignath & Buttner, 2008; Hacker et al., 2009; Hacker et al., 2000; Hadwin & Webster, 2013; Miller & Geraci, 2011; Nietfeld & Schraw, 2002; Zimmerman & Schunk, 2003) to improve calibration accuracy.

The majority of calibration studies have been conducted in laboratory settings under controlled conditions (Gutierrez, 2017; Gutierrez et al., 2017) and possess strong internal validity such as Boekaerts and Rozendaal (2010), Bol et al. (2012), Hadwin and Webster (2013), Higham (2013), Miller and Geraci (2011), Nietfeld et al. (2006b), Ozsoy (2012), Roderer and Roebers (2013), and van Loon et al. (2013). Nonetheless, some researchers, including Lundeberg and Fox (1991), McCormick (2003), Nelson and Narens (1994), Roderer and Roebers (2013), and Winne (2004), have argued that due to the complexities inherent in naturalistic learning environments, some of the lab findings may not be generalisable to classroom settings. To clarify some of the differences, Roderer and Roebers (2013) argue that in a lab setting, tests are designed by researchers strictly to elicit the maximum desired effect. Further, tests are administered by people unknown to students in an environment with no motivational relevance and minimum resemblance to a classroom.

On the one hand, the development of calibration in a natural setting such as a classroom is more complex than in a laboratory setting, with many forms of materials and teaching methods in use and a much longer time-gap between teaching and testing episodes; on the other, "80% of calibration studies are not situated in school subject-matter areas such as mathematics or science" (Parkinson et al., 2010, p. 11), as they focus on law-enforcement, medical diagnostic, and criminal justice matters. Some researchers (e.g., Parkinson et al., 2010; Jacobse & Harskamp, 2012) have warned that this issue makes it difficult to generalise the research findings to academic areas, given the domain specificity of calibration (Tobias & Everson, 1996; Veenman & Spaans, 2005). Moreover, calibration studies have almost exclusively used convenience samples of undergraduate students, which means that the developmental aspects of calibration for teenagers and young adults have not been captured in calibration research (Parkinson et al., 2010).

On a positive note, the current gaps in calibration research, as presented chronologically in Appendix B have the potential to underscore the future direction of calibration research and where the researchers might next focus their attention and efforts. Although the issues highlighted in Appendix B are limited to the topics pertinent to this study, further exploration of these issues may be beneficial as direction for future research on calibration.

Inconclusive or Contradictory Results in Calibration Research

As noted by Alexander and colleagues (2010), for a variety of reasons, such as the lack of clear conceptualisation (Schunk, 2008), calibration researchers' attempts to ascertain whether calibration accuracy can be improved have produced inconclusive and sometimes conflicting results. For instance, in a lab setting, feedback is assumed to improve calibration accuracy (Stone, 2000; Zimmerman, 1990). In a classroom setting, however, Nietfeld et al. (2005) could not replicate this result. To provide another example, in 2001, Flannelly demonstrated that in a classroom context, test difficulty affects calibration accuracy, while Bol and Hacker (2001) concluded that test difficulty does not affect calibration accuracy in a laboratory setting. The erroneous assumption that the results of the lab-based studies are necessarily generalisable to the complex setting of classrooms may be partly responsible for conflicting results. Further, calibration has often been identified as a purely psychological construct, which has been noted by some researchers (e.g., Bol et al., 2008a; Bol et al., 2005; Chen & Zimmerman, 2007; Hacker et al., 2008; Nietfeld et al., 2006a) as the reason why self-factors factors such as students' self-perceptions are largely absent from calibration research.

Contradictory or inconclusive results continue to trouble calibration researchers. Appendix C represents a concise list of the contradictory research results that are pertinent to this study. These issues include the conceptual grounding of calibration (#1 to #4), malleability of calibration (#5 to #8), reflection (#9 and #10), feedback (#10 to #14), calibration in a classroom (#15 to # 17), and confidence judgments (also called monitoring judgments, #18 and #19).

Rationale for This Study

"But I thought I knew this chapter inside out."

"I sort of had a feeling that I knew my stuff."

"I actually studied for this test and still got only 40%."

The above and similar comments might sound familiar to experienced mathematics teachers who realise early in their career that the accurate self-evaluation of task preparedness and performance is not a 'given' for students. Decades of research (e.g., Barnett & Hixon, 1997; Bol, Hacker and colleagues, 2000, 2001, 2005, 2008a, 2008b; Glenberg & Epstein, 1985; Kruger & Dunning, 1999; Flannelly, 2001; Nietfeld et al, 2005; Stoten, 2019; Winne, 2004; Winne & Jamison-Noel, 2002) have indicated that the inability to self-assess task performance can be detrimental for students' learning, motivation, and academic achievement. Through this study, I responded to the call for more research on students' calibration accuracy in a classroom context (e.g., Gutierrez et al., 2017; Huff & Nietfeld, 2009; Stoten, 2019), at K-12 level (McCormick, 2003; Winne, 2004), using mixed methods (Barnett & Hixon, 1997) of research.

My literature review generated for me fresh and constructive ideas about what kind of intervention might be effective in a mathematics classroom. Based on my teaching experience and due to the value assigned to explicit monitoring strategy training, feedback, and reflection in calibration research literature, I focused on incorporating these elements into my research design.

The result was a classroom intervention in which the elements of (a) monitoring strategy training, (b) curriculum-based classroom tests, (c) students' calibration graphs, and (d) students' reflection notecards merged with students' day-to-day learning activities and curriculum-based tasks. Inspired by what I learned from my literature review, as I noted earlier in Chapter One, the research question that I pursued was as follows: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom?

Chapter Two Summary

By reviewing the metacognition and calibration research literature, I intended to understand and present what researchers in this field had previously done, what they had not done, and how to position my study within the current scholarly discourse in calibration research in mathematics education. I began Chapter Two with a review of the historical roots and conceptualisation of metacognition and an examination of Nelson and Narens's (1990) theory of metacognition, which is the theoretical framework for my thesis. By attending to the social side of metacognition, I explained why I placed my study within Bandura's (1986) social cognitive theory and elaborated on the epistemological and ontological orientation of this study. I also reviewed the perspectives on metacognition and outlined the intervention studies in mathematics classrooms. I discussed the notion of calibration, its importance for students, and the difficulties inherent in conceptualisation and assessment of calibration. I also attended to the classroom interventions that attempted to improve students' calibration and identified the findings, the gaps, and the inconclusive or contradictory results in calibration research. I concluded Chapter Two by explaining my rationale for conducting this study and provided a chapter summary. In Chapter Three, I present the details of my research methodology and design for this study.

Chapter Three

Research Methodology and Design

Introduction

In this study, I used a pedagogical classroom intervention to explore the research question: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom? I applied Nelson and Narens's (1990) theory of metacognition within Bandura's (1986) social cognitive theory to study high school students' calibration in a Grade 11 mathematics course in which metacognitive strategy training, feedback, and reflection were embedded in curriculum-referenced instructions and curriculum-based classroom tests.

As I discussed in detail in Chapter Two, calibration refers to the accuracy of students' monitoring judgements. These judgements are students' interpretations of their own thinking processes and as such are in a constant state of flux, often not explicitly quantifiable. Moreover, some researchers (e.g., Garner & Alexander, 1989; Ozturk, 2017; Thomas, 2013; Veenman et al., 2006) suggest that metacognition is a multi-layered and complex phenomenon and propose that due to the complexities associated with a myriad of naturally occurring events in a classroom, conducting research on metacognition in a classroom setting may be problematic. As I noted earlier, employing purely quantitative methods of data collection and analysis might have been partly responsible for inconclusive or contradictory results in prior calibration research. Thomas (2013) stated that "if credible assertions regarding the efficacy of any intervention are to be made, it is appropriate to seek both qualitative and quantitative forms of data" (p. 6).

In their examination of 710 research articles published in six prominent journals between 1995 and 2005 in the field of mathematics education, Hart et al. (2009) found that 50% of the researchers used only qualitative methods such as observations, field notes, and interviews, 21%

used only quantitative methods such as descriptive statistics, correlations, and quasi-experiments, and 29% used mixed methods, which involves the use of both qualitative and quantitative methods, such as in this study. Some mathematics education researchers have raised concerns regarding the dominance of qualitative methods in mathematics education research. For instance, Silver (2004) asserts that in mathematics education, researchers have “erected a monument to qualitative research methods and non-experimental models of inquiry” (p. 154) during the past few decades, despite the increasing level of complexity of current educational issues that warrant multifaceted research designs. Facing the challenge of choosing a research methodology that could validly and accurately capture the dynamic nature of students’ calibration, I became convinced that the results obtained through a combination of both qualitative and quantitative methods could enrich my understanding of calibration and provoke constructive ideas to answer my research question, which demanded such an approach.

I begin Chapter Three with an examination of the mixed methods methodology that I applied in this study and introduce the research participants. Subsequently, I present the context and the research design, describe the intervention and data collection components, and highlight some quality concerns. I conclude Chapter Three by addressing the assumptions, limitations, and delimitations of my research design and providing a chapter summary.

Mixed Methods Methodology in Theory and Practice

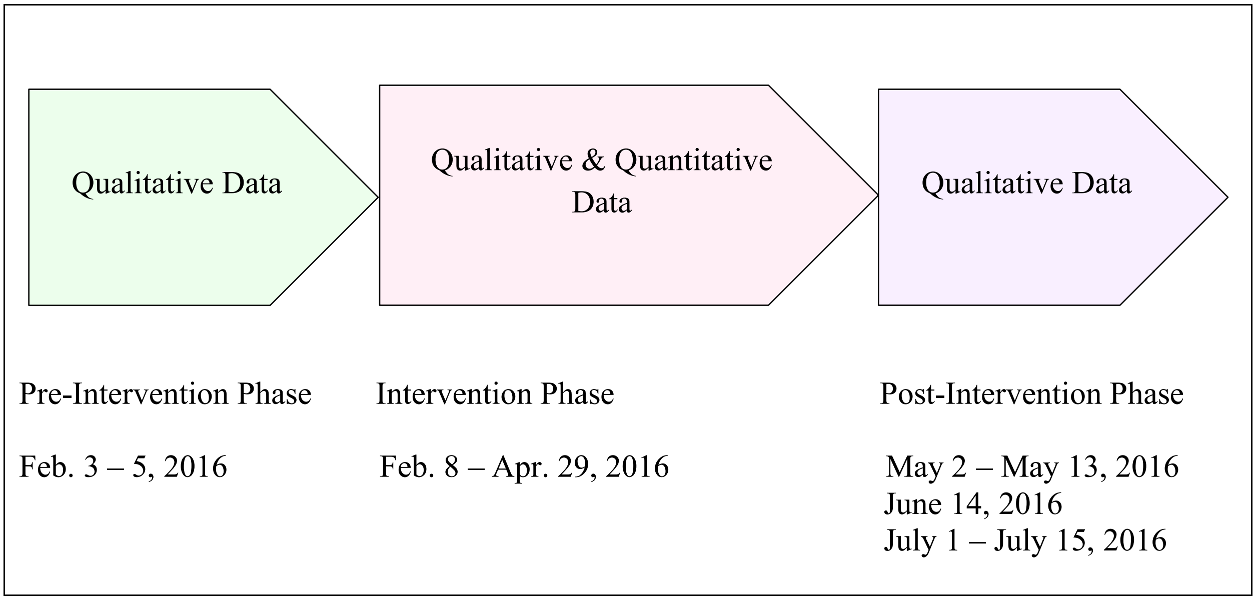
In this study, I used a mixed methods approach, which Teddlie and Tashakkori (2009) defined as including a combination of both qualitative and quantitative methods to best answer one or more research questions. For the purpose of this study, I adopted the definition of Hart et al. (2009), who broadly described a mixed methods approach as “the use of both qualitative and quantitative methods, in any part of the study” (p. 29). Hart et al. (2009) explained that they chose this definition to be inclusive without being constrained "by the issues of balance between qualitative and quantitative methods or integration in data collection, analysis, results, or conclusions" (p. 29). Hence, the current study qualifies as a mixed-method study of calibration because I used both qualitative and quantitative methods throughout the study.

Onwuegbuzie and colleagues (2009) propose that the following two factors can determine the mixed methods design:

1. Priority: According to this criterion, the researcher can give the same weight or different weights to the qualitative and quantitative aspects of the study. In this study, I gave equal weight and status to both aspects of the study, based on my understanding that the research question of this study could not be deeply and meaningfully answered by using either qualitative or quantitative methods alone, and that both methods substantially contributed to my understanding of students' calibration.
2. Implementation of data collection: This criterion refers to the order in which data are collected. The two methods of data collection are as follows: collecting data at the same time (i.e., concurrent design) or at different times (i.e., sequential design). Due to the iterative and dynamic nature of a high school classroom, my research design does not fully align with either style. Had I felt the necessity to label my study as sequential or concurrent, I would have called it a partially sequential mixed methods study of calibration. This is due to the sequential structure of the study, which included a pre-intervention phase, an intervention phase, and a post-intervention phase. Within the intervention phase, I collected both qualitative and quantitative data concurrently. Figure 3.1 depicts the sequential structure of my research.

Figure 3.1

Data Collection in Three Phases of This Research



In this study, each phase dialectically informed the next phase in a hermeneutic fashion. For instance, the extent of students' calibration awareness in the pre-intervention phase was important for my understanding of the reasons for students' miscalibration within the intervention phase. In choosing a mixed methods methodology for my study of calibration, I set five main goals. These goals for conducting mixed methods research were outlined by Greene, Caracelli, and Graham (1989) after they reviewed 57 mixed methods studies:

1. Triangulation: Seeking the convergence of results from different methods,
2. Complementarity: Seeking the enhancement and clarification of results from one method with the results from another method,
3. Development: Seeking to use the results from one method to help develop or inform the other method in terms of sampling, implementation, and measurement decisions,

4. Initiation: Seeking the discovery of paradoxes and contradictions, new perspectives, the casting of questions or results from one method with another; and
5. Expansion: Seeking to extend the breadth and range of inquiry using different methods for different inquiry components.

To further explain these goals, Johnson and Christensen (2008) stated that triangulation refers to the use of different pieces of evidence to explain a certain result with confidence, which improves the credibility and trustworthiness of the findings. Some scholars (e.g., Denscombe, 2010; Patton, 2002) have called for not only data triangulation by using different data sources, but also investigator triangulation (involving different researchers), methods triangulation (the use of varied methods), and theory triangulation (employing multiple theories) to interpret a single set of data. In this study, I sought methodological triangulation by collecting qualitative and quantitative data and by comparing and contrasting the results with the findings of the research literature in the field. Considering that the qualitative and quantitative methodologies are different but compatible (Johnson & Onwuegbuzie, 2004), I compared and contrasted the findings of these philosophical stands with what I could learn from calibration research literature.

I sought complementarity by overlapping the qualitative and quantitative data as I looked for consistency in emerging results and findings. I used the numerical values of each student's calibration at a particular point in time in conjunction with the same student's written comments at the same point in time. The key point was to compare and contrast the data from markedly different quantitative and qualitative approaches (Denscombe, 2010) and search this blended data for emerging ideas. Subsequently, I incorporated students' interview data into this blended data in pursuit of affirmation, accuracy, and convergence of the results.

I did not undertake development, in the sense of using the results from one method to develop the other methods in terms of sampling, implementation, and measurement decisions in this study. Based on my research design and the ethics considerations, I did not have access to the entire student data until after the semester ended and the follow-up interviews were completed. Although the results from the quantitative data informed the results of the qualitative data and vice versa, this was after the fact and did not affect any of my sampling or implementation decisions.

Initiation was sought through my attempts to find explanations for what appeared as paradoxes and contradictions in calibration research. For instance, why is calibration sometimes resistant to change? What is the nature of students' self-estimation judgments, or why may some students continue to miscalibrate despite feedback?

Finally, I approached the expansion of the breadth and range of my inquiry by using different data collection methods for different components of the inquiry. For instance, the methods used for the qualitative inquiry included reflection notecards, questionnaires, and follow-up interviews, while the methods used in the quantitative component were based on numerical and graphical assessments of calibration. Moreover, in my view, the examination of the development of students' calibration over time, in an actual high school classroom, using authentic curriculum-based material for research, enhanced the depth and breadth of this study by bringing the research context closer to the actual circumstances within which students' calibration might develop over time.

Research Participants

My inquiry into the change and development of mathematics students' calibration took me through the works of a significant number of researchers and scholars who have written extensively on the subject, with substantial knowledge, enthusiasm, and authority. Throughout my research of the topic, from Woodworth and Thorndike's (1900) paper on judgments of magnitude, similar to the notion of calibration, to Stoten's (2019) paper on calibration accuracy, and much in between, I often felt that I had personally met these scholars by following the evolution of their thoughts and by carefully studying and analysing their work over the span of several decades.

However, to my surprise, I always found a missing element in this vast body of knowledge: the participants. Who were these research participants who were involved in calibration research? Historically and philosophically, where were they coming from? What were their stories, goals, hopes, and aspirations? What were the foundations of their research responses and what were they confident about?

These questions inspired my decision to conduct qualitative research, allowing for the expressions of students' thoughts and stories in my work. In my view, the numbers, variables, graphs, written words, and spoken assertions that I have collected in my study have no intrinsic value in isolation from the person who contributed them. As such, I emphasise that in this study, the contributions were made by the researcher (myself) and the research participants (my students) who gave a new dimension to this study—the self.

I invited all students in the mathematics course that I taught (23 students) to participate voluntarily in this classroom research. They all consented and participated. Although students were aware that they could choose a different section of the same course within the host school without my research component, none of the students left the group. One of the students in another section of the course requested to join the research group; hence, I had 24 research participants. The students in the course were under no obligation to participate and could withdraw from the study at any time.

Research Setting, Context, and Environment

In this study, the host school was a public high school (grades 9-12) in a metropolitan school board in a large Canadian city with a multiculturally diverse population. The school had many visiting and exchange students on study permits, as well as immigrant and refugee students from all over the world. Concerning the country of origin of the student population at the host school, the last school census (Yau et al., 2018) reported that 64% of the students and 86% of their parents were born outside Canada and 70% of the students primarily use a language other than English at home. Consequently, language translators, social workers, case workers, volunteers, and guidance counsellors worked in this school.

To participate in this research, the School Board required each participant to provide a signed parent/guardian informed consent form; therefore, such consents were obtained. In addition, I asked the students for their signed assent form to ensure that I did not include their data in this research based on only parental consent and against their will (see Appendices D & E). Furthermore, I invited the students and their parents/guardians to attend a Research Information Night (see Appendix F). This provided an opportunity for me to meet and greet the students and their parents before the commencement of the study in order to provide context, offer first-hand information, and address their questions and concerns.

The students providing data for this study were in a grade 11, university-bound mathematics course, which I taught during the second semester (February–June) of 2016. Before the commencement of the study, the research design and procedures received the necessary ethics approvals from the University of Alberta Research Ethics Board (REB 2) and the District

School Board's External Research Review Committee (ERRC), as presented in Appendices M and N.

As noted earlier, although developing metacognitive skills has a well-documented positive influence on students' academic achievements in mathematics, the present study was a study of the development of students' calibration in the context of classroom teaching, using mathematics as a subject context within which to conduct the research. As such, the focus of this study was on neither the cognitive aspects of mathematics learning nor on improving task performance outcomes (i.e., students' grades). In other words, the quality of students' grades in the course and whether these grades increased or decreased had no bearing on my study; I strictly focused on the discrepancy between students estimated and achieved performance outcomes (i.e., their calibration accuracy) and how this discrepancy changed over time.

Moreover, this study was a non-experimental exploration of students' calibration and although the number of research participants was small (24 students), conducting research with similarly small number of participants is not uncommon in calibration research or in metacognition research in general. For instance, in 2009, Huff and Nietfeld conducted calibration research with 24 participants. Roderer and Roebers' (2013) calibration research involved 19 students, and Gutierrez and colleagues (2017) conducted mixed methods research on calibration involving nine students. Thomas and McRobbie (2001) engaged 24 students in their metacognition research, and Thomas (2014) conducted metacognition research that included 33 grade 11 students. For researchers with a purely quantitative experimental viewpoint, the small number of participants might be a major dilemma because quantitative researchers are mainly concerned with the generalizability of their research findings. For researchers who adhere to a non-experimental standpoint, such as myself, however, the number of research participants is not of utmost concern due to the prominence given to the meaning inherent in people's lived experiences (Denscombe, 2010). Therefore, for this mixed methods study of calibration, the small sample of 24 participants was reasonable.

The Importance of Learning Environments

Social cognitive theory (Bandura, 1986) places a high value on students' learning environments as one of the triadic (i.e., personal, behavioural, and environmental) factors affecting human learning and development. Thomas (2003) conducted research involving 1,026 high school students and defined seven main characteristics of his model of the Metacognitively Oriented Learning Environments, which I call *Thomas's model* in this thesis, for brief. The seven dimensions of Thomas's model are as follows: Metacognitive Demands, Student-Student Discourse, Student-Teacher Discourse, Students' Voice, Distributed Control, Teacher Encouragement and Support, and Emotional Support. In this section, I explain how the seven dimensions of this model provided me with a framework to establish a metacognitively oriented classroom environment for students, and to examine the extent to which the classroom environment in this study was conducive to the potential development of students' calibration.

Metacognitive Demands. Thomas (2003) suggested that in metacognitively oriented classroom environments, teachers place metacognitive demands on students, requiring their thoughtful engagement and participation in learning tasks. This might not be easy if students are used to being rewarded for passively attending the class and mindlessly attending to academic tasks. For instance, Thomas proposed that in a metacognitively oriented environment, in addition to teaching effective metacognitive strategies students should be asked to trial such thinking strategies and have opportunities to evaluate their own metacognitive efficacy. In this study, teaching and practising metacognitive strategies were blended with students' daily problem-solving activities. Calibration sessions were enacted and reinforced by me with the students' help and participation. Moreover, by completing the calibration graphs I directed students' attention to the extent of the accuracy of their calibration and by completing the reflection notecards students had the opportunity to reflect on how effectively they could self-evaluate.

Student-Student Discourse. Thomas (2003) recommends student-to-student discourse. I encouraged a classroom culture in which students' questions and conversations about calibration were given credence, and student to student discourse was recognised as a necessary and valuable part of the daily routines and classroom activities. For example, students were in conversation with each other in their daily group-work sessions during which I taught them monitoring strategies. Further, they exchanged ideas in completing their calibration graphs and worked with each other to help me enact ideas during the weekly calibration sessions (see Appendix I for scripts).

Student-Teacher Discourse. Thomas (2003) emphasises the importance of student-teacher discourse. My calibration sessions encouraged student involvement and the daily monitoring strategy training procedures were meant as opportunities for students to ask questions of the teacher about calibration and to share their thoughts and comments with me.

As I began teaching the students calibration sessions and monitoring strategies, I noticed that some students were hesitant about speaking up, sharing views, and asking questions about calibration. I assumed that this was due to the lack of experience, on students' part, in terms of verbalising their thoughts and views about calibration. As time passed by, conversations became more frequent and students began to use the terminology that I used in teaching them about calibration.

Students' Voice. I frequently reminded the students that their questions and comments regarding any aspect of my classroom intervention were welcome and appreciated. Over the course of the intervention, questions frequently arose about why we were talking about metacognition and calibration in a mathematics classroom, how monitoring strategies were helpful in mathematical problem solving, and why calibration graphs were a necessary aspect of my intervention. By addressing these and similar questions in detail and also by enlisting students' direct participation during my enactment of the weekly calibration sessions, gradually, a safe space opened up in which students realized that it was not only allowed, but also encouraged to question any aspect of my pedagogical plans and methods. Since these communications, questions, and concerns were expressions of students' thinking about their own thinking processes (e.g., why do I need to monitor my problem-solving steps?), they could potentially contribute to the development of students' metacognition (Thomas, 2003). Further, by conducting follow-up interviews, I had the opportunity to incorporate students' voices and views directly into my research.

Distributed Control. Thomas (2003) suggests that having the appropriate tools of communication such as the appropriate language facility and terminology is essential to the development of students' metacognition. With calibration likely constituting a new concept for students, as a first step in sharing control with students, I worked toward creating a sense of familiarity with the necessary language and terminology to use in class when talking about calibration.

On all curriculum-based classroom test, students had the opportunity to choose their own test questions from a selection of questions that I provided. This style of distributed control encourages students to use their metacognition and choose the questions that they believe they can best answer, which further supports the development of their calibration considering that one aspect of students' calibration is to distinguish between what they can or cannot do on a test. For example, given the choice, a student may decide to choose a graphing question rather than a question that requires algebra.

Teacher Encouragement and Support. Thomas (2003) argues that the classroom teacher's support and encouragement has a positive influence on the development of students' metacognition. In my study, due to the diversity of the host school, students in the language learning program needed a great deal of one-on-one support to deal with the daily classroom activities. I emphasised on and modelled an inclusive classroom atmosphere, a *togetherness* in which everybody was encouraged to support others and was praised for it.

Emotional Support. A supportive classroom culture in which students feel welcome, cared for, and supported is in line with the views (Bandura, 1986; Efklides, 2011; Thomas, 2003) that consider affective and support issues to be integral to the development of students' metacognition. In the context of the current study, emotional support for all the students was consistently forthcoming throughout the daily learning activities, research related tasks, and individual and group assessment protocols. For instance, students had the opportunity to work in pairs or groups, projects and presentations were peer evaluated, and students were able to choose their test questions.

Overall, the dimensions of the model of Metacognitively Oriented Learning Environments helped me to pursue a classroom environment in which it was possible to develop and enhance students' metacognition. Table 3.1 describes the seven dimensions of the model of Metacognitively Oriented Learning Environments (Thomas, 2003) with examples of each dimension in a mathematics classroom context.

Table 3.1

Description of the Model of the Metacognitively Oriented Learning Environments Dimensions and an Example of Each Dimension

Dimension name	Description (Extent to which:)	Example (In this mathematics classroom:)
1. Metacognitive demands	Students are asked to be aware of their metacognition and to develop their calibration.	Students were asked the *What*, *Why*, & *How* questions by their teacher while solving problems.
2. Student-student discourse	Students discuss their calibration with each other.	Students engaged in pair and group discussions during the calibration sessions.
3. Student-teacher discourse	Students discuss their calibration with their teacher.	Students discussed their calibration graphs and their *What, Why*, & *How* strategies with their teacher.
4. Student voice	Students feel it is legitimate to question the teacher's pedagogical plans and methods.	It was OK for students to ask the teacher why, for example, they did calibration graphs.
5. Distributed control	Students collaborate with the teacher to plan their learning activities as they develop as autonomous learners.	Students decided which test questions they wanted to answer from a large pool of test questions available to them.
6. Teacher encouragement and support	Students are encouraged by the teacher to improve their calibration accuracy.	The teacher encouraged, guided, and supported students who sought ways to improve their calibration accuracy.
7. Emotional support	Students are cared for emotionally in relation to the development of their calibration.	Students' opinions and concerns were addressed, discussed, and respected by their teacher.

Teaching Weekly Calibration Sessions

To cultivate a metacognitively oriented learning environment, to enrich students' metacognitive knowledge, and to provide context and terminology for students' conversations with me and with each other, I conducted 10 weekly calibration sessions, each one 10 to 15 minutes long, throughout the intervention phase. In each session, I first taught a brief calibration lesson and then provided scenarios and examples to model the concept. Appendix I provides the full scripts that I prepared for my 10 calibration sessions and explains my ways of modelling of the concepts for students. Each of these sessions encompassed an important aspect of calibration relevant to its classroom applications. For example, in the first session, which was focused on the importance of self-evaluation, I started working on a word-problem on the board and students were carrying out their own solutions in groups. I made some errors but did not correct myself and continued to solve the problem. In the end, I wrote a final statement, gave myself a 90% estimation, and stopped working. Students commented that because I had a 90% estimation of my work in mind, I did not feel the necessity to Check and Reflect—a monitoring strategy that I had introduced to the students, earlier in the semester.

Further, I intended to help students construct metacognitive knowledge that they needed to enhance their monitoring processes. As I detailed earlier in this thesis, Nelson and Narens's (1990) theory suggests that through the enrichment of monitoring processes, students develop an awareness of the possible control actions they might take to attain their intended goals. Against this theoretical backdrop, my calibration sessions were intended to promote control actions and support the dimensions of Thomas's model for a metacognitively oriented classroom environment (Thomas, 2003), while working on the triadic influences of person (e.g., building confidence), behaviour (e.g., improving study habits), and environment (e.g., providing an inclusive and supporting atmosphere), in line with Bandura's (1986) social cognitive theory. Building on the social dimension of metacognition, calibration sessions were inclusive and encouraged the cooperation and participation of all students. Moreover, while teaching and modelling calibration sessions, I chose to use the relevant terms such as metacognition (instead of, for instance, reflection), calibration, monitoring, and other relevant terms to build concept knowledge and terminology and facilitate future references and conversations. The dates and titles of the calibration sessions were as follows:

1. Session #1 (February 9, 2016): The self-evaluation of task performance matters.
2. Session #2 (February 16, 2016): What does it mean to be calibrated?
3. Session #3 (February 23, 2016): Overestimation and underestimation of performance
4. Sessions #4 (March 1, 2016): What are the estimation judgments made of?
5. Session #5 (March 8, 2016): The critical role of overestimation and underestimation in guiding study behaviour
6. Session #6 (March 22, 2016): I understand calibration, but what is metacognition?
7. Sessions #7 (March 29, 2016): Metacognition and mathematical problem-solving
8. Session #8 (April 5, 2016): Importance of calibration graphs
9. Session #9 (April 12, 2016): Thinking fast and slow on mathematical tasks
10. Session #10 (April 19, 2016): Where to next?

Prior to this research, as a mathematics teacher, my classroom conversations with students often centred around mathematical thinking and algorithmic competencies. Calibration sessions afforded me the opportunity to highlight for students the importance of being a good judge of those competencies. In hindsight, I propose that the calibration sessions likely added to students' classroom experiences at a personal level (e.g., students were taught about their own cognition and metacognition), a behavioural level (e.g., students were reminded about their test preparation habits and study behaviour), and an environmental level (e.g., classroom setting and atmosphere were metacognitively supportive for students), consistent with the social cognitive theory and the dimensions of the model of Metacognitively Oriented Learning Environments (Thomas, 2003).

The Problem of the Assessment of Calibration

An ongoing debate exists among calibration researchers (e.g., Allwood, Jonsson, & Granhag, 2005; Boekaerts & Rozendaal, 2010; Dunn, 2004; Keren, 1991; Koriat & Levy-Sadot, 2000; Nelson, 1996; Nietfeld et al., 2006a; Rutherford, 2017a; Schraw, 1995; Schraw et al., 1993) regarding how to best assess calibration. Calibration refers to the accuracy of students' monitoring judgements; therefore, the debate centres on how these judgements can be assessed accurately and reliably.

Schraw (2009) suggested that the accurate and reliable assessment of calibration is built on four basic dimensions of scale, timing, granularity, and type of assessment. The choice of scale depends on the research design. Within the body of calibration research, scholars have assessed monitoring judgements on a continuous scale (e.g., ranging from 0% to 100%; Schraw, 2009) or a non-continuous scale (e.g., the dichotomous prediction of successful versus unsuccessful performance; Nietfeld et al., 2006a).

Schraw (2009) explains that 'timing' refers to whether monitoring judgements are elicited before the completion of a task (i.e., predictions), during the task (i.e., concurrent), or after the completion of a task (i.e., postdictions), while 'granularity' describes whether the judgements are made about a specific item (i.e., local judgements) or an entire task (i.e., global judgements). Finally, the type of assessment (Schraw, 2009) affects the conclusions drawn. For instance, if a student makes a global monitoring judgement of 80% on an exam and achieves 85%, the researcher can calculate the absolute difference between estimation and performance outcomes called calibration (Bol &Hacker, 2001; Huff & Nietfeld, 2009) of + 5%.

Further, the researcher may decide to compute the directional difference between the estimation and performance which, to borrow from the calibration research terminology, is called 'bias' (Chen, 2002; Mengelkamp & Bannert, 2010) of –5%, which indicates a small measure of negative bias or slight underestimation.

In my study, the students made their monitoring judgments on a continuous scale of 0% to 100%, following each test performance episode on the entire test, providing global judgements. From a quantitative standpoint, I calculated both absolute accuracy (i.e., calibration) and bias (i.e., over-/underestimation) to gain insights into both absolute and directional accuracy of students' monitoring judgments.

In the assessment of calibration, the choices of scale, timing, granularity, and type of assessment are critical decisions that require the careful examination of the theoretical underpinnings of the study as well as the intended research question. Because I based this study on a classroom intervention, I opted for calculating the global absolute accuracy and bias of students' postdictive monitoring judgements, which according to Schraw (2009) is the recommended method in the assessment of monitoring accuracy when the effectiveness of an intervention or training is under examination.

Research Design

Several considerations guided my research design. Nelson and Narens's (1990) theory of metacognition posits that monitoring drives control processes. This implies that enhanced monitoring may possibly lead to better control decisions. For instance, enhanced monitoring may result in students contemplating taking notes while reading. As such, I designed the first two components of my classroom intervention—namely, monitoring strategy training and curriculum-based classroom tests—to support and develop students' monitoring processes.

The classroom monitoring strategy training focused on improving students' monitoring skills by encouraging students to ask themselves questions such as "What am I being asked to do?", "Am I using the right method?", "Why is this the right method?", or "Are my answers correct and meaningful?" while engaged in various mathematical tasks. On the other hand, the classroom tests were intended to develop students' monitoring skills by allowing them to choose their own test questions from a list of available questions on each test. Here, the intention was to encourage students to think carefully and distinguish between what they thought they knew and what they did not know yet or did not know well.

The social cognitive theory of Bandura (1986) places a high value on both the social and personal aspects of metacognition with regards to students' learning and development. Hence, the other two components of my classroom intervention—feedback and reflection—were designed to support the development of students' calibration by promoting the social and personal aspects of metacognition in alignment with the social cognitive theory.

Overall, I based the four components of my intervention on Nelson and Narens's (1990) theory of metacognition and Bandura's (1986) social cognitive theory, within the metacognitively oriented learning environment of a mathematics classroom.

The Centrality of Calibration Change to my Study

Essential to my study was the exploration of change in students' calibration, both conceptually and numerically. As I stated earlier, some calibration researchers propose that calibration is malleable and can be developed (e.g., Gutierrez de Blume, 2017), while others (e.g., Bol, Hacker, O'Shea, & Allen, 2005) posit that calibration is resistant to change. In my research design, I employed a mixed methods approach which included both the qualitative methods such as reflection notecards to explore *why* students' calibration changed, as well as the quantitative methods such as the performance estimations to explore *how* and to what extent the measures of calibration accuracy changed over time.

In the pre-intervention phase, my focus was on gaining insights into students' metacognitive knowledge in terms of the extent of their awareness of their own calibration accuracy. In this phase, students completed an open-ended questionnaire (i.e., Questionnaire #1, see Appendix G) in which they answered four questions.

In the intervention phase I undertook to support the development of students' metacognitive knowledge through monitoring strategy training sessions and classroom conversations. Further, I sought to help students develop their calibration and narrow the gap between their estimated and achieved performance outcomes by teaching weekly calibration sessions. In this phase, students wrote weekly tests, estimated, and graphed their performance scores, and completed their individual reflection notecards after each test.

Finally, in the post-intervention phase, I explored how students' expressions of metacognitive knowledge and their calibration had changed over time. In this phase, students completed another open-ended questionnaire (i.e., Questionnaire #2, see Appendix H) and 12 students participated in the follow-up interviews.

The end result of my research design was a teacher-developed, curriculum-embedded, three-phase intervention that I implemented to develop students' calibration, by using monitoring strategy training, curriculum-based classroom tests, students' calibration graphs, and students' reflection notecards. Table 3.2 provides the mixed methods data collection protocol that I used in this study.

Table 3.2

Mixed Methods Data Collection Protocol

Pre-Intervention Phase	**Intervention Phase**	**Post-Intervention Phase**
Qualitative Data (q)	Qualitative (q) & Quantitative (#) Data	Qualitative Data (q)
Methods/Instruments	Methods/Instruments	Methods/Instruments
Questionnaire #1	Monitoring accuracy training strategies, curriculum-based tests, self-estimations of performance, calibration graphs, reflection notecards	Questionnaire #2 and Follow-up student interviews
Data Collection	Data Collection	Data Collection
Students' written responses as a baseline for their awareness of the extent of their calibration (q)	Test performance (#), self-estimations of performance (#), calibration graphs (#), and reflection notecards (q)	Students' written responses relating their calibration awareness and reasons for miscalibration (q) Students' verbal responses regarding their experiences/ theories of calibration (q)

Data Collection Components and Procedures

As Presented in Table 3.2, I collected both qualitative and quantitative data in my study. Except for the follow-up interview data, my qualitative data consisted of two questionnaires completed by the students, one at the beginning and one in the end of the intervention phase, and students' weekly reflection notecards. After being completed by the students, I collected the questionnaires and the notecards by sending large sealable envelopes around the classroom. The envelopes were then sealed and sent to the main office to be kept in the vault.

The quantitative data consisted of students' weekly test performance outcomes, their test performance estimations, and students' calibration graphs. As the classroom teacher, I had access to students' performance outcomes. However, due to research ethics considerations, I did not collect or keep any record of students' performance estimations. A copy of students' weekly tests which included students' test performance outcomes and test performance estimations was sent to the main office to be kept in the vault as research data. Students' individual calibration graph templates were also kept in the vault, but once a week, after each test was returned to the students, I arranged for the envelope that contained the graph templates to be brought into the classroom by a school office administrator so that the students could update their calibration graphs. Students updated their graphs by point plotting their weekly test performance outcome and test performance estimation on their individual graph templates. With ethics permission, students made their last performance estimation on their final exam on June 14, 2016. On June 16, 2016, which was school's exam review day, students reviewed their exams, completed their last reflection notecard, and updated their calibration graphs for the last time. Appendix J represents a concise version of the calendar that I used as the timeline for the intervention-related classroom activities between February 3, 2016 and May 13, 2016. The last day of the semester was June 30, 2016.

The last element in my data collection procedures took place during the first two weeks of July 2016. During this period, I conducted the follow-up interviews which were only audio-recorded. After transcription, the audio files for these interviews were stored in an encrypted, password-protected digital folder in the main office, in the host school.

Data Collection in the Pre-Intervention Phase

In this initial phase of my pedagogical intervention, the instrument that I used was Student Questionnaire #1. I presented this questionnaire to the students on February 8, 2016, and again on February 10 and 11, for those who were either new to the course or had missed it the first time. Questionnaire #1 consisted of the following four questions:

1. *Are you sometimes surprised by your score when you get a test or quiz back? Explain***.**

   By asking this question, I intended to direct the students' attention to a possible discrepancy between their expected and achieved performance outcomes, identifying a conceptual baseline for calibration awareness before the commencement of the classroom intervention.

2. *If sometimes there is a gap between the score you expected and the one you received, in your view, what are the reasons for this gap***?**

   Here, the intention was to help the students develop a preliminary conceptualisation of calibration and for me, to gain insights into the students' possible reasons for miscalibration. For instance, I was curious to explore whether these reasons for miscalibration were internally oriented (e.g., I did not study for this test), externally oriented (e.g., this was a tricky test), or both.

3. *After writing a mathematics test, do you think you can accurately estimate your test score? If so, how do you think you will make this estimation, and what will you base this estimation on?*

   The question of whether students could accurately estimate their performance outcomes was foundational to my study and was incorporated into my methodology and design. This question was meant to bring the concept of performance estimation into focus for students. It was also an early exploration of the nature of students' performance judgements.

4. *Do you usually receive a grade that is more than you expected, less than you expected, or the same as you expected? Why do you think this is the case?*

This question was an introduction to the concept of directionality of miscalibration (also called bias) and was raised to encourage the students to think about the overestimation and underestimation of their performance outcomes.

The students' pre-intervention responses to the above noted questions on Questionnaire #1, at the beginning of the study, were of immense importance to me, as the responses spoke to the extent of the students' initial awareness of their calibration accuracy and its directionality (i.e., over- or underestimation of performance outcomes). The questions were open-ended, and students were encouraged to write extensively at will and include any comments they considered appropriate. I collected the questionnaires by sending a large sealable envelope around the classroom. The envelope was then sealed and sent to the main office where it was kept in the vault until the end of the school year. I accessed the questionnaires after the follow-up interviews were completed.

Four Components of my Classroom Intervention

Monitoring Strategy Training

Monitoring strategies are essential to learning and task performance (Butler & Winne, 1995; Winne, 2001) and in Glaser and Chi's (1988) view, they are critical to the development of human expertise in general. Although calibration accuracy among students is not a certainty (Glenberg & Epstein, 1985; Pressley, Ghatala, Woloshyn, & Pirie, 1990; Sheldrake, 2016; Stoten, 2019), to date, few researchers have attempted to devise effective classroom interventions to integrate monitoring training strategies into students' daily learning activities, which might be considered surprising given that metacognitive skills are deemed malleable.

To devise monitoring strategies that work in a mathematics classroom setting, while attending to providing a metacognitively oriented atmosphere for students, I built a classroom protocol based on my teaching experiences. This protocol took into account the appropriate use of monitoring strategies in a mathematics classroom as well as the insights that Polya (1973) and Schoenfeld (1992) offered on how to redirect students' attention to their monitoring processes. For instance, while solving an equation, my goal was to ensure that in addition to using the usual cognitive processes, such as mathematical algorithms, students could evaluate their own progress. As such, I informed the students that I might ask them the following questions while they were engaged in mathematical tasks:

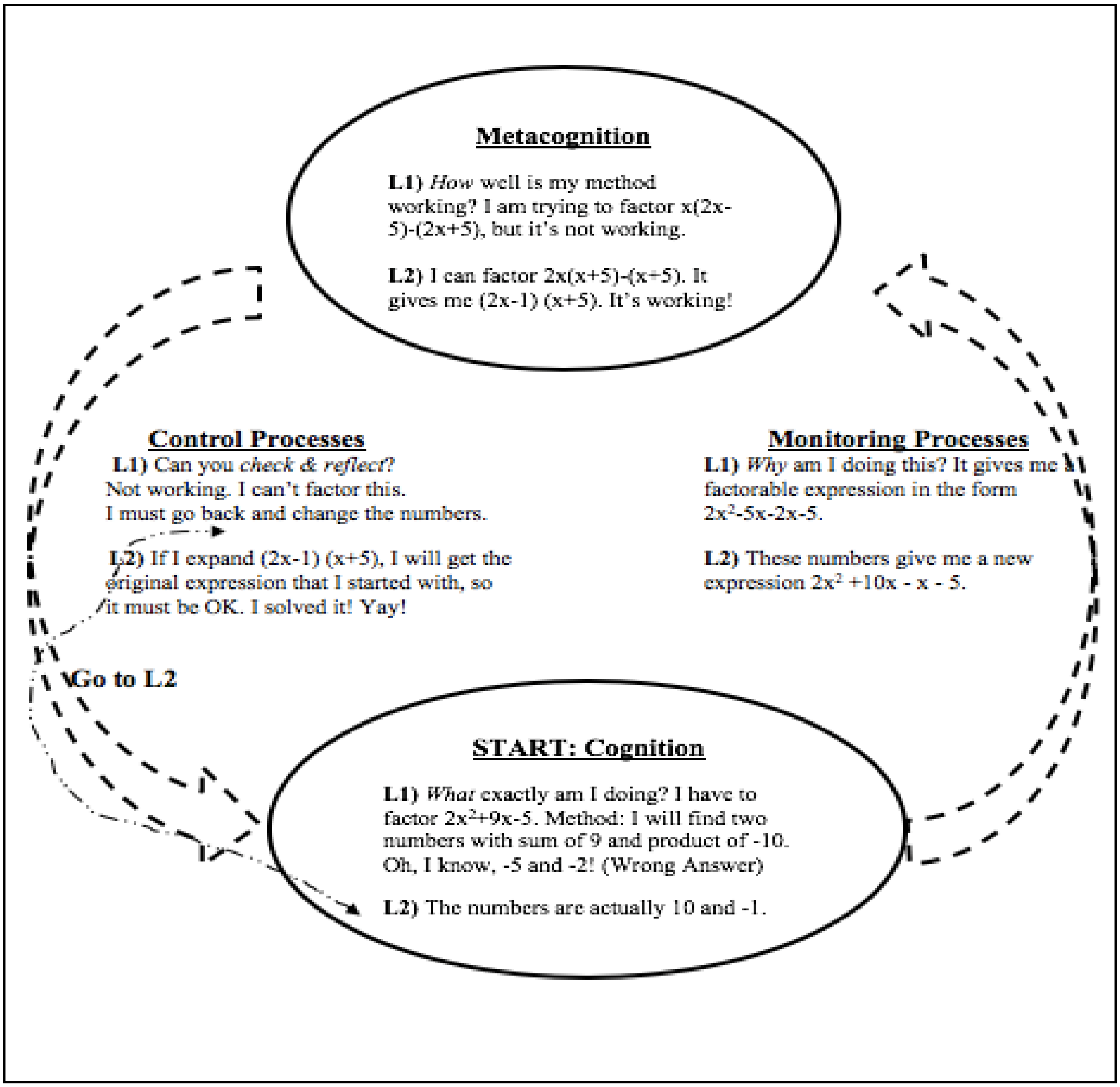
1. Can you describe *what* exactly you are doing?
2. *Why* are you doing this (i.e., why do you think it fits into the solution)?
3. *How* well do you think your method is working so far?
4. Can you *check* and *reflect* to see if your answers are correct and meaningful?

These questions are consistent with Nelson and Narens's (1990) theory of metacognition, in which the information (i.e., metacognitive knowledge) obtained through effective monitoring enhances the control processes, allowing students to reconsider ineffective and inaccurate mathematical procedures. In choosing these questions, my intention was to help further develop students' metacognitive knowledge through monitoring and I was hopeful that all students, even those who had not yet achieved procedural fluency and could not complete any mathematical procedure without experiencing several unsuccessful attempts might benefit from monitoring their own thinking, through the stages of mathematical tasks.

Nelson and Narens's (1990) theory of metacognition further envisions a cyclic process in which many levels or rounds of monitoring and control may take place. To explicate the importance of accurate monitoring (i.e., good calibration) in guiding students' control decisions while engaged in mathematical tasks, Figure 3.2 illustrates two levels of monitoring and control, which I called L1 and L2. At the start of the first level, Monitoring L1 informs Metacognition L1. Subsequently, Metacognition L1 triggers Control L1 and the first level ends. However, at the end of the first level, the squiggly dashed line on the left side of Figure 3.2 represents the transition from the first level to the second level which includes Monitoring L2, Metacognition L2, and Control L2.

Figure 3.2

A Classroom Example of Two Levels of Monitoring and Control Processes



To explain what Figure 3.2 communicates, suppose that while working on factoring trinomials in algebra, as I go around the classroom, I ask the students to think about *what* they are exactly doing, *why* they are doing it, *how* their method is working, and that they *check* their work *and reflect* (i.e., emphasising monitoring strategies). One student in the group, Shirin (a pseudonym), thinks about the task before her: *what* am I exactly doing? At this level, Cognition Level 1, she uses the method that she was taught, and searches for two key numbers to factor the trinomial. Suppose she finds the two numbers, but her answer is incorrect. Without monitoring, Shirin might go ahead without correcting her answer. I continue by asking her *why* she is doing this (Monitoring Level 1), and *how* well her method is working (Metacognition Level 1). Encouraged to think about *why* and *how*, Shirin notices that the expansion of the factored form does not produce the initial algebraic expression. Next, I insist that she should *check* her answer and *reflect* on its accuracy. Recognising that her answer is incorrect, she may decide to go back and adjust (Control Level 1) her solution. Shirin is now at Cognition Level 2. She works out a new pair of numbers, monitors her work (Monitoring Level 2), and double checks (Control Level 2) and realises that the new set of numbers are producing an expression identical to what she started with. She is now ready to write down her accurate final answer.

Curriculum-Based Classroom Tests

The second component of my research design relied on the curriculum-based tests that the students wrote in the course. In total, students wrote 12 *tests* in the course. Six of these were quizzes, which included 6 questions and four were unit tests, which included 10 questions. Students also wrote a midterm exam (15 questions) and a final exam (30 questions). Although I use the term ‘test’ for all 12, in my quantitative analysis, I make a distinction by calling a quiz, *a small task*, while calling the unit tests, the midterm exam, and the final exam *large tasks.* To be sure, students' summative and formative assessment and evaluation in the course included many other opportunities such as projects, individual and group assignments, mathematics laboratory reports, and presentations. However, consistent with calibration research and to keep the volume of the data manageable I only included the data that I collected by utilising these 12 curriculum-based classroom tests in my research data.

Johnson and Allwood (2003) cited that research on calibration has often associated the calibration on one task with the calibration on an unrelated task either locally (i.e., on one item) or globally (i.e., on the entire task). For instance, when the link between calibration and students' performance is measured based on the students' GPAs, the researchers may associate higher GPA with better calibration without being able to posit a theory about why and how calibration was developed in the first place. Did calibration improve because students with higher GPAs made their test estimations based on their previous performance outcomes on similar or different academic subjects, changed their study habits, or used monitoring strategies? Further, when conducting calibration research, the same students may not be available to participate across related and meaningful tasks over an extended period. In this study, I reduced unsystematic interferences by engaging with the same students over an entire semester.

Additionally, examining students' calibration in one session (e.g., Chen, 2002; Koku & Qureshi, 2004) is misaligned with a social cognitive view of learning and development (Bandura, 1986; Zimmerman, 1989) in which as a metacognitive phenomenon, calibration changes incrementally not only within the individual, but also based on the interactions of the individual with others and the environment. I was hence convinced that studying calibration over time, using classroom tests, measuring calibration on and at the same time as the classroom tests, and simultaneously collecting students' comments about their calibration would best suit my inquiry.

To embed a part of my intervention in classroom tests, I employed a method that, to my knowledge, has not been previously used in calibration research in a high school mathematics classroom. This method involved adding an extra number of questions to all the curriculum-based classroom tests. For example, on the first test, I offered the students 12 questions, but instructed them to select and answer only six questions that they thought they could answer most accurately. I advised the students not to attempt more than six questions due to time constraints, but more importantly because the goal was to develop students' monitoring skills in distinguishing between what they thought they knew and did not know as of yet, which is an aspect of students' calibration.

In my review of calibration research literature, the only study using a similar method of assessment was Isaacson and Fujita's (2006) work with undergraduate educational psychology students. However, in Isaacson and Fujita's (2006) study, the tests included only true/false and multiple-choice questions, and the students' estimations were made before their task performance. Hence, in Isaacson and Fujita's study, the students' estimations were predictions of task performance, whereas in my study, high school students worked on full-solution questions, engaged in problem-solving, and made their estimations after their task performance in the form of postdictions of task performance. As noted earlier, calibration researchers suggest that students' postdictions of task performance are more accurate because postdictions are based on richer metacognitive knowledge.

In my experience, mathematics teachers often adhere to strict assessment and evaluation procedures, hoping to achieve a fair assessment for all students. One might argue that a fair assessment involves the assessment of what students know, not what they do not know. To further explain, consider the following scenario. As noted earlier, all students' names are pseudonyms: A trigonometry quiz has five questions. Katie who has learned the concepts and practised with similar questions receives a perfect score. Aryana who has practised with other questions and knows some other concepts not included in the quiz can correctly answer only two questions. Aryana fails the quiz and decides that she is ‘not a math person’ and that she ‘hates math’. On the second quiz, however, the teacher provides ten questions and asks the students to choose and answer five. Both Katie and Aryana receive perfect scores, feel like ‘math persons,’ and neither of them ‘hates math.’

In my study, after returning each test to the students, I took up all the questions in detail with students’ help and participation. Moreover, after each test, I posted a full and detailed solution set to our Google Classroom for the benefit of those students who needed assistance on questions they had avoided on the test. By providing this example, I intend to demonstrate that different styles of assessment (e.g., having or not having a choice in terms of answering the test questions) may pose different demands on students’ monitoring. In the above-noted example the second style of assessment, which is the style I used in my study, encourages careful monitoring, and supports students’ motivation and achievement.

In using this assessment approach, I pursued two goals. First, as Polya (1973) and Schoenfeld (1992) noted, most mathematics students rush into a solution attempt without thinking much as soon as they encounter a question, only to realise later that they do not really know how to solve the question. By giving the students the opportunity to choose their own test questions, I intended to engage their monitoring skills before attempting a solution. Second, I hoped that by operationalising monitoring as a self-controlled action, the students might realise their own autonomy in successful task performance. Figure 3.3 represents an example of the 12 classroom tests that I used as intervention.

Figure 3.3

An Example of an Assessment Task Used as Intervention

MCR3U Quiz #1_Equations ....../ 12 = ...%

Friday, Feb. 19/ 2016

Name__________________

My estimated percentage mark is

..........%

Instructions:

1) *There are 12 questions on this quiz. Please read them all very carefully and* ***choose and answer only 6 questions*** *that you know best how to answer. Each question is worth 2 marks with part-marks given for partial work.*
2) ***After*** *you finished, please estimate your mark and write it in the circle. Thank you!*

---

1. Without solving, explain how you think you might solve $2y^2 - 10 = 0$, for *y*?
2. Without solving, explain how you think you might solve $4 + p = \sqrt{p + 6}$, for *p*?
3. Solve $x^2 - 2xy + 1 = y$, for y.
4. Give an example of an equation that can be solved by using the Zero-Product-Property. Do not solve your equation.
5. Solve for *t*, $2t^2 - 61 = 101$.
6. Solve for *m*, $\frac{4}{m+3} = \frac{m-8}{9-m^2}$.
7. What do you think it means to "solve an equation"?
8. Solve $x^2 - 5x - 36 = 0$.
9. Solve $2y = \sqrt{21y - 5}$.
10. What do you think an "inadmissible value" means for an equation?
11. Write an example of an equation, which has "5" and "– 2" as its roots (or solutions).
12. Solve for *x*, $\frac{2}{x^2} - \frac{1}{x} = 0$.

Attention:

In these six circles, please indicate which 6 questions you want to be marked:

On classroom tests, the questions varied in terms of how they made the inquiry. To explain, based on the provincial Ministry of Education guidelines (Ontario Ministry of Education, 2010), formative and summative evaluation tasks are written and evaluated in four achievement categories of Knowledge and Understanding, Application, Communication, and Thinking. For instance, in Figure 3.3, question #3 is considered an application-type question in that it combines different mathematical concepts such as re-arranging, factoring, and solving in the form of a mathematical algorithm, whereas question #10 is considered a communication-type question in which students are expected to explain their mathematical thinking using full and clear statements. These guidelines are consistently followed in Ontario across all grade levels and subject areas and were observed throughout the semester; therefore, the categories (i.e., the different types of questions on a test) were not considered as defining factors affecting students' calibration.

Students' Calibration Graphs

The third component of my research design incorporated both metacognitive feedback and performance feedback as intervention to develop the students' calibration. Feedback is often conceptualised as the information provided by an external agent, such as a teacher, regarding the aspects of one's performance. Hattie and Timperley (2007) suggested that feedback may operationalise students' efforts to reduce the discrepancy between their current and desired performance (i.e., improved calibration) by increasing effort and developing effective error-detection strategies.

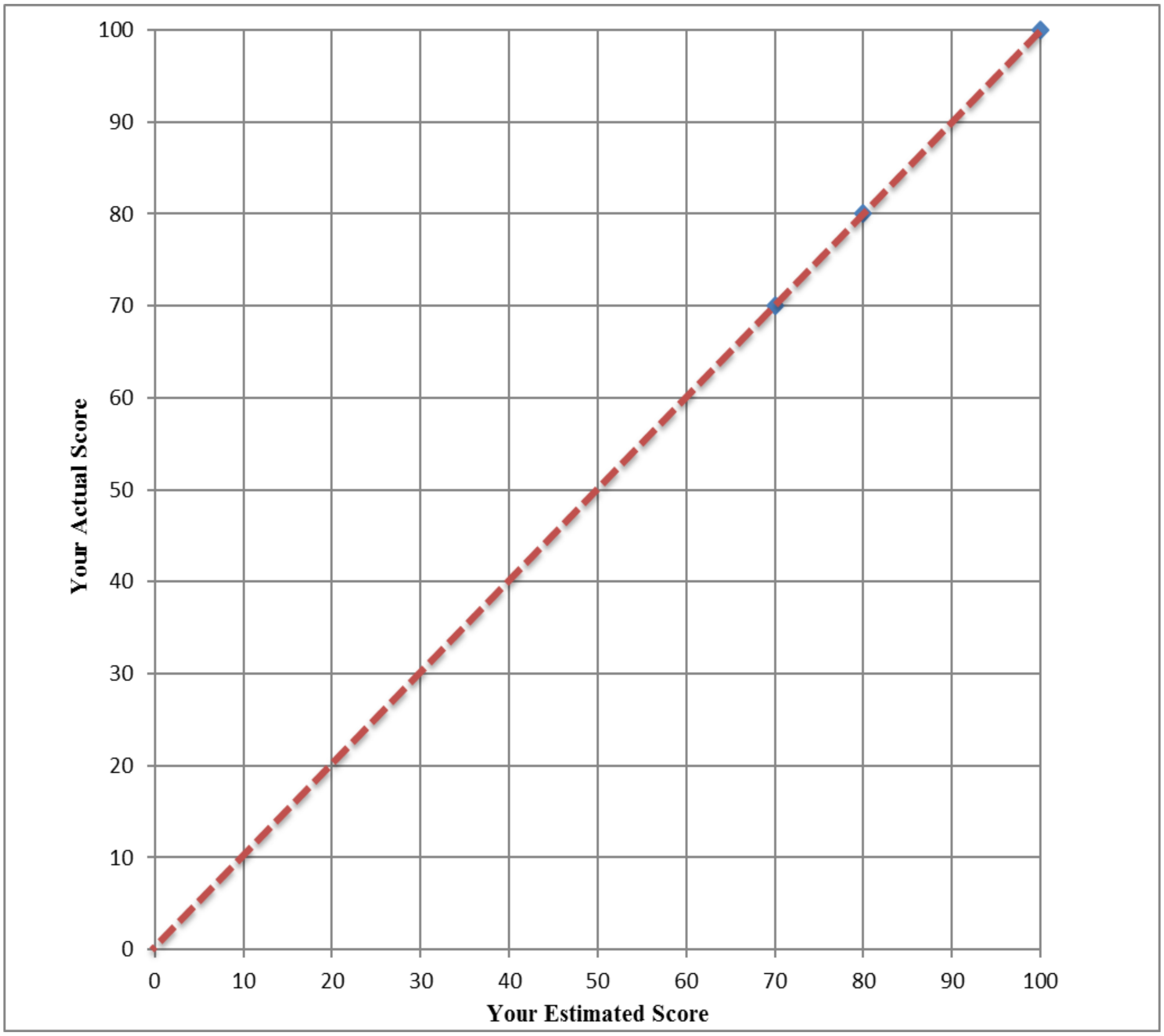
According to Hattie and Timperley (2007), calibration graphs intend to inspire three questions: First, *where am I going?* While completing their calibration graphs, the students could see on their own calibration graph whether they were approaching the line of perfect calibration (the diagonal line; see Figure 3.4) or moving away from it. Second, *how am I doing?* The visual clarity and simplicity of calibration graphs was intended to help students observe and monitor how successful they were in closing the gap between their perceptions of performance outcomes and their actual performance outcomes. Third, *where to next?* Theoretically, to the extent that students are calibrated, they can make informed decisions (Labuhn et al., 2010; Winne, 1995) and plan their future time allocation and learning activities.

As a framework for feedback, calibration graphs work at all four levels of Hattie and Timperley's (2007) theory on effective feedback. To explain, calibration graphs work at (a) the task level, providing students with external feedback about each test, (b) the process level, illustrating the change in one's performance and calibration over time, (c) the self-organisation level, visually tracking and representing one's estimations versus calibration accuracy, and (d) the person level, working as personal evaluation tools to provide an optimistic forward-looking outlook that places the control and ownership of performance in the learner's hand.

In the present study, the students received calibration graph templates that they updated after each weekly test, with their own performance estimations and actual performance outcomes. These graphs provided the students with visual feedback regarding their weekly performance outcomes (i.e., performance feedback) and the accuracy of their calibration (i.e., metacognitive feedback). After completion, I collected the calibration graphs in large envelopes and sent the sealed envelopes to the main office where the graphs were kept in the vault. I accessed the calibration graphs and all my qualitative data (questionnaires and notecards) after the completion of the follow-up interviews, two weeks after the end of the school year. Figure 3.4 illustrates a blank calibration template with the diagonal line representing the line of perfect calibration, where the performance estimation matches the actual performance outcome.

Figure 3.4

A Blank Calibration Graph Illustrating the Line of Perfect Calibration



Students' Reflection Notecards

The fourth and final component of my research design involved reflection. In Nelson and Narens's (1990) theory of metacognition, monitoring is perceived as an information-driven dimension of metacognition in a sense that its activity is compelled by information (i.e., metacognitive knowledge; van Overschelde, 2008), and not by other factors such as personal will or motivation. One possible interpretation suggests that to the extent that students possess metacognitive knowledge about their monitoring processes, they will be able to make accurate judgments regarding their performance. Hence, by developing students' metacognitive knowledge, reflection may further support monitoring processes leading to improved calibration.

As a classroom teacher, prior to this study, I had noticed that whenever I returned the graded tests to students, some students did not spend much time to reflect on their work and my corrections and comments. Hence, as an element in my research, I introduced a student self-recorded *reflection notecard* to be completed after each test. On each reflection notecard, students answered the following two questions: 1) Why do you think there is a gap between the two grades (i.e., the estimated and achieved), and 2) What strategies are you planning to use to close the gap? By asking these two questions, my goal was to direct students' attention to not only their performance outcomes but also to the *gap* between their performance estimations and their actual performance outcomes. In some studies, self-reflection has been used as a single independent method for improving students' metacognition. For instance, Huff and Nietfeld (2009) conducted an experimental study in which as a single method of intervention, the student's attention was directed to the discrepancy between their performance estimations and performance outcomes to encourage a habit of good calibration. The researchers reported that both treatment groups gained in confidence and calibration accuracy.

As outlined by Schoenfeld (1992), some social cognitive theories of learning posit that learners develop most of their thinking skills through socialisation, rather than direct instruction. Owing to the completion of the reflection notecards as a classroom activity, and due to my emphasis on the student-student discourse dimension of the model of Metacognitively Oriented Learning Environments (Thomas, 2003) used in my study, some social comparison (Labuhn et al., 2010) among the students was expected; it was likely that the students would become aware of the approximate levels of their classmates' academic task performance or their calibration accuracy compared to their own. Although in complex social interactions within a classroom setting, social comparison may be inevitable (France-Kaatrude & Smith, 1985), not all social comparisons necessarily have detrimental effects. In my study, I did not attempt to influence, limit, or control student-student discourse, which was, in my view, essential to the natural development of students' metacognition. In fact, the resulting socialisation among the students was an aspect of my attempts to create and support a metacognitively oriented learning environment for students. As I noted earlier (see Table 3.1), student-student discourse is the second of the seven dimensions necessary to establish a metacognitively oriented classroom environment for students (Thomas, 2003) and essential for the development of students' metacognition.

Based on my research design and ethics considerations, I did not have access to my entire data while conducting my study. Consequently, I was not aware of the students' comments on their reflection notecards. Moreover, before the commencement of my study, the students in the course were advised to return their signed consent and assent forms to the main office at the host school, not to me. Hence, throughout the study I did not know who the actual research participants were, who had submitted their signed consent forms to the main office. I accessed my entire research data, after the follow-up interviews were completed, two weeks after the end of the school year. Figure 3.5 represents a blank reflection notecard that students completed in my study.

Figure 3.5

A Blank Reflection Notecard

Student Notecard – Quiz #1

Name____________________ Date___________

My estimated grade was ....

My actual grade was ....

1) Why do you think there is a gap between these two grades (if there is)?

2) What strategies are you planning to use to close the gap?

Data Collection in the Intervention Phase

In this phase, I collected both qualitative and quantitative data concurrently. The instruments used in this phase were as follows:

Weekly Measures of Test Performance Outcomes

The test scores used in this study were the measures of the actual curriculum-based classroom tests that the students wrote over the course of the semester. To ensure authenticity, the tests were strictly content based, in a format familiar to students, prepared by the classroom teacher (i.e., myself). Based on the guidelines of the provincial Ministry of Education (Ontario Ministry of Education, 2010), these tests were designed within the achievement categories of Knowledge and Understanding, Application, Communication, and Thinking. In total, students wrote 12 tests and had the opportunity to choose their test questions from a list of potential test questions that I made available to them. Further, I used the same continuous scale of 0% to 100% for grading and recorded the results as percentage points ranging from 0% to 100%.

Weekly Measures of Performance Outcome Estimations

On all tests, the students were instructed to first answer the test questions that they had selected from a list of potential test questions that I had made available to them and then make a performance estimation of their expected performance outcome as a percentage between 0% and 100%. Students were not given any instructions as to how to base their estimations. For instance, they were not instructed to base their performance estimations on their previous test results, the test content, or their hours of study. Calibration research (Desoete & Roeyers, 2006; Roderer & Roebers, 2013) suggests that these estimations are ‘self-evaluations’ of task performance because they are made after the completion of a task and they are more accurate than students’ predictions of their task performance.

Weekly Measures of Calibration Accuracy

In line with the previous studies on calibration (e.g., Roeder & Roebers, 2013), the absolute value of the numerical difference between each performance outcome and its corresponding performance estimation was computed as a measure of calibration accuracy. For instance, if a student made an estimation of 95% but achieved 85%, the measure of calibration would be 10%, which indicates the absolute degree of deviation between task performance estimation and the actual task performance outcome. Being computed as an absolute value, a calibration measure is represented as a positive number, which for instance in the above-mentioned case does not distinguish between an overestimation of 10% and an underestimation of 10%. Hence, in this study, the bias (i.e., measure of overestimation or underestimation) scores were also computed to gain insights into the directionality of calibration. In the above-noted example, the student's performance demonstrates a 10% positive bias (i.e., overestimation).

Weekly Reflection Notecards

Studies on metacognition have depicted a complex picture of the interplay between self-reflection and mathematical tasks. In an earlier study by Schoenfeld (1985), the author concluded that without reflection, mathematically knowledgeable students may perform poorly on mathematical tasks. This was important to this study because the quantitative measures of calibration are inherently linked to students' performance on mathematical tasks and without reflection, these performance outcomes might not always present a realistic picture of students' mathematical performance. Ramdass and Zimmerman (2008) reported that more reflective students develop a deeper and more coherent understanding of the course content than their less reflective peers leading to improved monitoring accuracy.

In this study, students completed individual reflection notecards every week after receiving their graded tests. After completion, the notecards were collected and sent to the main office, where they were kept in the vault. Each reflection notecard recorded the student's estimated and actual performance scores and asked the following two questions: *(a) Why do you think there is a gap between these two grades (if there is)?* and *(b) If such a gap exists, what strategies are you planning to use to close the gap?* Here, the intention was to direct students' attention to the gap between their estimated and achieved performance outcomes and to encourage them to contemplate strategies that might improve their calibration.

Weekly Completion of the Calibration Graphs

During the intervention phase, students completed their personal calibration graphs once a week. Each point on a calibration graph corresponds with two quantitative measures: the estimated score of a test on the x-axis and the actual test score on the y-axis (see Figure 3.4). For a variety of reasons, the data collected on calibration graphs were integral to my study. To explain, these graphs were completed at the specific points in time simultaneous with students' comments on their reflection notecards; these were crucial points at which both the qualitative and quantitative data were simultaneously collected and were available for each individual student.

Further, as noted earlier, feedback was one of the four components of my intervention and the calibration graphs construed both cognitive grade feedback provided by me, as well as metacognitive calibration feedback provided by students for students. The data collected on calibration graphs were foundational to my understanding of how students' calibration changed over time.

Data Collection in the Post-Intervention Phase

The post-intervention phase was the last stage of my data collection. In this phase the students completed another questionnaire (i.e., Student Questionnaire #2, see Appendix H) and I conducted the follow-up interviews with 12 students who had volunteered to be interviewed. Questionnaire #2 was presented to all the students on May 2, 2016, and was repeated on May 4, 2016, for those who had missed it the first time. I conducted the student interviews during the first 2 weeks of July 2016. The qualitative data collected in the post-intervention phase focused on the extent of students' calibration awareness and their conceptual understanding of the notion of calibration compared to the initial conceptual baseline, before the intervention phase began.

Student Questionnaire #2

This open-ended questionnaire included the following questions:

1. *You have been participating in educational research for the past 14 weeks. Congratulations! What does calibration mean to you? What do you think it does?*

   By asking this question in the post-intervention phase, I intended to explore the students' personal theories of calibration, the development of their metacognitive knowledge, and how they made sense of the concept.

2. *In your experience, how did your calibration change and develop over time, if it did?*

   The concept of change was central to my study. The quantitative data could determine the numerical change in students' calibration. However, I felt compelled to ask this question to explore the conceptual change in students' awareness of their own calibration.

3. *Each time that you noticed a gap between your estimated and actual scores, what did you think were the main reasons for the gap?*

Why do students overestimate or underestimate their performance? In addition to the insights from calibration research, I intended to learn the answer to this question directly from the students.

4. *Each time you made a performance estimation, how and based on what did you make your estimation?*

   As discussed earlier in Chapter Two, the bases of students' performance estimations are not quite known to calibration researchers. I decided to explore the topic by directly asking the students how and based on what their performance estimations were made.

Within the theoretical framework of this study, the two questionnaires asked similar questions at the beginning and in the end of the intervention phase so that the conceptual change in the students' awareness of calibration could be explored qualitatively.

Students' Voices: Follow-Up Interviews

My deep interest in students' personal theories of calibration made me wonder how the students made sense of learning about the presumably novel notion of calibration, and what they thought calibration was or could do. These wonderings inspired the follow-up interviews within the post-intervention phase of this study. The interview questions were not restrictive, as each question was meant to initiate a conversation. Although it was possible to answer each question briefly, the interviewees often described their views in detail, thereby adding valuable depth and insights to the study. The interview questions were as follows:

1. *Why, in your opinion, there might be a gap between students' perceptions of performance and their actual performance in mathematics?*

I had already asked the students this question on their reflection notecards. However, in the post intervention phase, I repeated the question, looking for clarification and confirmation of students' previously stated views.

2. *In what way did you personally try to close this gap, and did your method work for you?*

   This was also a reflection notecard question. However, in this post-intervention phase, I intended to explore if students had benefitted from my intervention and if they had experienced any success in closing the gap between their estimated and achieved performance outcomes.

3. *We humans are powerful theory builders, aren't we? What is your personal theory of calibration, and how do you make sense of it?*

   As a researcher, whether and how students made sense of the concept of calibration was immensely important to me. This question was my attempt to explore the extent of students' sense making about their own calibration in the post intervention phase.

4. *Reflecting on your experience in this course, did your calibration change over the course of the semester? Can you explain?*

   The numerical change in each student's calibration was something that I could determine based on my quantitative data. However, those numbers could not reveal to me whether or not the students were aware of their calibration change.

Interviews are often considered essential to conducting qualitative research because, as Johnson and Christensen (2008) noted, they provide rich information about participants' thoughts, ideas, beliefs, knowledge, reasoning, motivation, and feelings about the topic under study. Before the commencement of the interviews, I had collected both qualitative and quantitative data to explore students' calibration change both numerically and conceptually. As such, student interviews were my last chance to seek clarification, confirmation, and convergence of the qualitative and quantitative results. In conducting face-to-face interviews, I felt compelled to deepen my understanding of the students' perspectives on calibration and the ways in which they made sense of this newly met concept, which might ultimately lead to concept internalisation and behavioural change. In essence, these interviews served as a window of opportunity for an open exchange of ideas among the research participants – me and the students. As the School Board required, I conducted the interviews after the semester ended, during the first 2 weeks of July 2016 after the grades had been assigned. All 24 students were invited to participate in the follow-up interviews, of which 12 students agreed to participate. The follow-up interviews took place in the mathematics office at the host school, and each interview was approximately 20 minutes long. As I started to review the follow-up interview data, I realised that a detailed and systematic analysis was required to reveal the essence of what was brought to light during these interviews. Upon reflection, I decided to save the rich qualitative data of the follow-up interviews to be used as the last stage of my analysis to check for clarification, confirmation, and convergence of the results.

Research Quality in Mixed Methods Design

Johnson and Christensen (2008) contend that in mixed methods research, researchers combine complementary strengths and non-overlapping weaknesses of both qualitative and quantitative methodologies to provide the best possible explanations for complex research questions. Alas, combining methods adds to the complexities associated with the evaluation of mixed methods research.

As I discussed in Chapter Two, calibration research is most often conducted experimentally in controlled settings, which—despite producing insights—might not bridge the gap between the laboratory findings and classroom practices. Neither might it inspire classroom teachers to conduct authentic research with students to connect theory and practice. To this end, I designed each of the methods used in this classroom study of calibration to meet the following criteria: (a) interested mathematics teachers could incorporate these methods in their daily practices with minimum disruption to students' learning activities, and (b) the methods would assist in the construction and collection of both qualitative and quantitative data. Considering the complexities associated with the evaluation of mixed methods research, in the next section, I address reliability and validity with regards to the quantitative strand, trustworthiness in relation to the qualitative strand, and the ethical research design in relation to both qualitative and quantitative strands of this thesis.

Reliability and Validity in the Quantitative Strand

To ascertain the reliability of the instruments used in quantitative research is essential to the quality of any quantitative research. In this study, in the intervention phase, I used the students' test performance outcomes and performance estimations to elicit the measures of calibration and bias, which mirrored the procedures used in the related studies of calibration, including Bol et al. (2005, 2010), Hacker et al. (2000), Nietfeld et al. (2005, 2006a), Roderer and Roebers (2013), Rutherford (2017a), and Stoten (2019). Many types of reliability measures have already been highlighted in research. For instance, Nunnally (1972) suggested three methods to ensure test reliability: (a) two alternative forms of the same test are used for the same group of students, and the scores are correlated, (b) the same test is used in two different occasions with different students, and the scores are correlated, and (c) the test is used once; then it is divided into two sections, and the scores for the two sections are correlated. Of the three methods, Nunnally (1972) considered the first method to be the most comprehensive. Having the advantage of being the classroom teacher, I used the first method. In so doing, I wrote two different versions of the first test in the course. Students wrote both versions of Test #1; after writing each test, they estimated their performance outcome scores. I collected students' grades and their estimations on both versions of Test #1. For each version of the first test, the absolute values of the difference between the students' estimations and achieved scores (i.e., their calibration measures) were computed. To explore the association between the students' calibration measures on the first and second versions of Test #1, as Nunnally recommended, I used a bivariable correlation analysis in which the calibration measures obtained from the first version of Test #1 were correlated with the calibration measures obtained from the second version of the same test using IBM SPSS 24 software. The results provided a Pearson correlation coefficient of 0.846, and the correlation was declared significant at 0.01 level ($\rho < .01$), which suggested an adequate level of reliability for the first test. Although the content variability among the different tests could not be avoided, owing to the consistency in the design, format, and difficulty level of the weekly classroom tests and to save valuable class time (i.e., not to ask the students to write two versions of every test, every week) the procedure was not repeated for the other tests. Although this issue speaks to the ecological validity of my study, since mathematics students usually do not write each of their tests twice, it may constitute a limitation for my study. My results were consistent with O'Connor's (1989) review of calibration research, suggesting that the test–retest reliability of the calibration measures was relatively high with the reliability coefficient ranging between 0.72 and 0.85.

Goodman (2008) proposes that the validity of a test refers to whether a test serves its purpose by showing what it claims to show, elaborating on four common forms of validity: construct validity, internal validity, external validity, and ecological validity. According to

Goodman (2008), construct validity refers to the extent to which a data collection instrument measures what it intends to measure. Internal validity is concerned with whether the resultant effects are caused by the intended factor. External validity deals with the generalisability of the sample to a wider population than the sample. Finally, ecological validity focuses on whether the context of data collection resembles the world.

In this study, I pursued construct validity (Goodman, 2008) through using curriculum-based tests, with a familiar and relevant content, and in a format expected by students. The testing sessions occurred 11 times throughout the classroom intervention and once for the final exam, measuring students' postdictions of performance outcomes as well as their actual performance outcomes, as was intended. Moreover, on all tests including the final exam, the students could choose and answer the test questions they felt confident about, thus supporting the construct validity of the tests as the data collection instruments that measured what they intended to measure—the students' performance outcomes on the mathematical tasks that they could actually perform.

In conducting this research, blending the intended intervention procedures as smoothly as possible with the students' daily activities depended upon maintaining the natural setting of the classroom, which included too many environmental variables to control as would be the case in experimental studies. For example, while completing their notecards or calibration graphs, students led conversations with me and with each other. In my view, this was an aspect of the natural and metacognitively oriented setting (student-student discourse and student-teacher discourse dimensions of the model of Metacognitively Oriented Learning Environments, Thomas, 2003) that I intended for this research. As such, I did not try to stop or redirect these conversations. Moreover, the research methods and procedures merged with the classroom tasks and activities for the sake of ecological validity (Goodman, 2008) of this study. To provide an example, I did not impose any limitations on what students could base their performance estimations on. For instance, I did not instruct the students to estimate their performance strictly based on their previous performance outcomes in the course. Students were free to base their estimations on anything they saw fit (e.g., hours of study, particular content such as trigonometry, overall test difficulty). As such, the internal validity of students' performance estimation data that I collected could not be substantiated. However, this lack of control possibly sustained the external validity (i.e., the context of data collection generalisable to a wider population than the participating students) and ecological validity (i.e., the context of data collection resembling a real classroom, not a laboratory) of my quantitative data.

Trustworthiness and an Ethical Approach to Research Design

Denzin and Lincoln (2005) elaborated that qualitative researchers face the inevitable *problem of representation*, as establishing a direct link between text and experience is problematic. Furthermore, Denzin and Lincoln argued that the interpretations of qualitative data are often evaluated in the light of the qualitative researchers' descriptions and interpretations as far as these interpretations influence the credibility, transferability, confirmability, dependability, and authenticity of research. Lincoln and Guba (1986) suggested that in qualitative research, credibility, transferability, dependability, and confirmability could be used respectively, as analogous for the internal validity, external validity, reliability, and objectivity used as the evaluation criteria in quantitative research.

In this study, I sought credibility (Guion, Diehl, & McDonald, 2002), which is analogous to internal validity in quantitative research, through the prolonged engagement of the participants with the construct of calibration which was the topic my inquiry. Further, as Guion et al. (2002)

recommended, I sought credibility through the cross-checking of the qualitative and quantitative findings with the research literature. Transferability, akin to external validity, was sustained through the ecologically valid (Goodman, 2008) conditions of the study in which the test contents were familiar and meaningful for the students, the students were invested in the outcome, and the context resembled the world. Authenticity in the sense of the data representing a balanced and truthful view of the participants' values and beliefs was achieved through the use of qualitative methods of inquiry, including offering the students open-ended questionnaires, reflection notecards, and follow-up interviews. Further, the research findings reflected authentic insights and views of the students (Bryman, 2001).

This brings me to an important aspect of this research—my students' well-being. Hostetler (2005) contended that any research in education bears moral and ethical implications, and it is up to the researchers whether they give voice to these implications or remain silent about them. In my dual role as a teacher–researcher, I found myself wondering whether this research was *good* for my students. Were the participating students often happy and engaged, or distant and resistant? What was the cost to these young people of becoming more metacognitive and self-evaluative than they already were?

In Chapter Four, I share what my students thought about this study, and in Chapter Five, I elaborate on the lessons I learned from my students. It is worth noting, however, that as a classroom teacher, I found that conducting research opened another dimension into how my students and I related to each other. It allowed for a *togetherness* and a different kind of communication that enriched our perceptions of each other's lives and realities. I doubt that we would have experienced this connection and togetherness had I remained just their *mathematics teacher*. I also think that in addition to this connection, it was what I call the 'ethical research design' that allowed for these other dimensions of human relations to unfold. Had I chosen an experimental research design, I would not have *met* my students in the novel and profound ways that I did.

Assumptions, Limitations, and Delimitations of my Research Design

Conducting scholarly research on metacognition in a mathematics classroom made me realise how critically restricted I was in my ways of seeking answers to inherently complex questions. I experienced limitations in the availability of resources, such as language translators (for parents), educational assistants, and modified computer labs for language learners in a mathematics course, but more importantly, I came to understand the complexity and multidimensionality of the matters at hand and recognised the assumptions, limitations, and delimitations in my thinking and decision making in conducting classroom research. In this section, I discuss the underlying assumptions, limitations, and delimitations of my study.

Underlying Assumptions of my Research Design

In conducting research, the three tenets that I firmly held were as follows: (a) With very few exceptions, all mathematics students, regardless of their current level of academic task performance, have the potential to perform better if taught effective metacognitive strategies; (b) In teaching mathematics, classroom teachers should pay attention to what students are confident about because students' cognitive activities and strategic study behaviour are influenced by these judgements; and (c) It takes more than purely cognitive procedures to learn mathematics deeply and meaningfully. Metacognitive strategies are particularly important to mathematical modelling, pattern recognition, and problem-solving.

In addition, I made the following three assumptions. First, in my perspective, mathematics education is a pillar of social justice. Mathematical literacy has been referred to as the key to the 21st century citizenship (Moses & Cobb, 2001). Schoenfeld (2009b) asserted that high school graduates' mathematics proficiency levels are considered predictors of university attendance and a fair chance at a dependable career. As such, I assumed that mathematics education will remain important to school programs. Second, I assumed that the same conceptual framework guides students' metacognitive judgements in various mathematics areas (e.g., algebra, geometry, trigonometry), mainly due to the absence of research addressing this issue. Finally, my third assumption was that based on the findings of current calibration research (Nietfeld et al., 2006b) the measures of students' calibration are possibly normally distributed. To explain, in their exploration of the distribution properties of calibration measures, Nietfeld, Enders, and Schraw (2006b) studied 10,000 cases. Their study suggested that the measures of calibration accuracy followed the bell-shaped curve of a normal distribution (Field, 2009) in which the majority of the largest calibration measures were situated around the centre of the curve, implying low levels of calibration accuracy for the majority of the cases.

Limitations of my Research Design

Considering limitations as those characteristics of my research that were out of my control (Simon, 2011), I perceive my epistemic beliefs about the nature of mathematical thinking as limitations that could potentially both strengthen and weaken my study in one way or another. I am cognizant that for me mathematical epistemology and mathematical pedagogy are profoundly intertwined. Thus, I acknowledge that the ways in which I conceptualise and present mathematical topics to my students may influence the features of classroom discourse and affect my students. Maggioni and Parkinson (2008) posited that by affecting classroom instructions, teachers' epistemological beliefs may play a role in students' learning and academic task performance.

By embracing social cognitive theory, I concur with Schoenfeld (1985) in viewing mathematics learning as an inherently social activity as well as a cognitive pursuit. Further, as mathematics learning involves familiarity, connectedness, conscious awareness, and logical thinking and understanding, in my pedagogy, I introduce mathematics not as mathematical information, but rather as *mathematical knowing*—a flexible, constantly evolving, personally and socially constructed way to study and interpret phenomena and to understand the world. I appreciate, however, that my view of the nature of mathematical inquiry invites a certain pedagogy often marked by the exploration of ideas, formulation of conjectures, and answering open-ended questions, rather than a predominant concern with *getting the correct answer*. I realise that this aspect of my pedagogical practice can occasionally frustrate some students who are not used to answering open-ended questions and forming conjectures in a mathematics course.

The limitations of this study also stemmed from the time constraints imposed by conducting the study concurrent with delivering the curriculum and from my attempts to minimise the disruptions to the students' classroom activities. Further limitations emerged from aligning my study schedules and protocols with the school's formative and summative assessment and evaluation policies, mathematics department's internal guidelines, and from students' ongoing involvement in a wide range of extracurricular activities.

Since my present study was embedded in a classroom intervention, the sample size, criteria, and the timeline for the intervention were out of my control. Lastly, as noted before, due to the ethics considerations in my research design, I had no access to my entire data throughout the intervention, which denied me the opportunity to provide ongoing adaptive data-based feedback for students. Although my adherence to these guidelines limited my scope, it added to the external and ecological validity of my study.

Delimitations of my Research Design

As Simon (2011) noted, delimitations are those characteristics of research that limit its scope and define its boundaries, but they are in the researcher's control. In this study, the choice of calibration as my research topic was the first delimitation. I chose calibration as a research topic because of my many years of observing mathematics students' frustration and dismay due to their miscalibration on mathematical tasks, and because I had witnessed how students' miscalibration could profoundly affect their self-confidence, motivation, future career decisions, and in a sense, their overall well-being. I also decided to study calibration due to my understanding that becoming reflective and self-evaluative matters to students beyond their high school career. Although choosing calibration as a research topic was a personal decision, it delimited the scope of my research.

Another delimitation was the choice of a mixed methods methodology, as I came to realise that some important aspects of the dynamic and complex nature of calibration were not quantifiable. On reflection, situating the study within the social cognitive theory of Bandura (1986) produced another delimitation, as the social cognitive theory framed and underpinned all the elements of the methodology and design. For instance, in alignment with the social cognitive theory (Bandura, 1986), the monitoring training strategies and calibration sessions were not only explicitly taught, but also modelled. Further, the students were offered many opportunities to discuss and practice the process of how to develop their calibration so that over time, the notion of being calibrated could become internalised and eventually habitual.

As I detailed in Chapter Two, Nelson and Narens's (1990) theory of metacognition suited the purpose of the study; however, choosing this theory over and above the other theories of metacognition constituted a theoretical delimitation for the conceptualisation of students' knowledge monitoring processes and how these processes influenced students' self-evaluative decisions in this study. To provide an example, Nelson and Narens's theory perceives calibration as a psychological cognitive phenomenon, and as such, many other aspects of students' learning (e.g., personal, motivational, and affective factors) are inconsequential in this theory. Further, the theory suggests a linear causal relationship between improved metacognitive monitoring and enhanced calibration. Situating this study within the social cognitive theory of Bandura (1986) and collecting and analysing qualitative data brought in students' voices, values, emotions, ideas, beliefs, knowledge, reasoning, and motivation (Christensen, 2008), thereby adding to the depth and breadth of the present study.

Overall Ethical Views and Benefits for Students

I understand and acknowledge the ethical concerns associated with classroom teachers doing research with students in their own classrooms. This is often a conundrum, given that classroom teachers are encouraged to engage in authentic classroom research. Many researchers (e.g., Hacker et al., 2000; Hacker et al., 2008b; McCormick, 2003; Nietfeld & Schraw, 2002; Nietfeld et al., 2006a; Parkinson et al., 2010; Zimmerman & Schunk, 2003) have noted that such studies on calibration are certainly in short supply. To date, I have not been able to identify any calibration studies performed in high school mathematics classrooms using a mixed methods methodology similar to this study with the explicit goal of enhancing students' calibration.

Potentially, conflict may arise from the dual role of a teacher–researcher. I disclosed the potential conflict of interest to the University of Alberta Research Ethics Board (REB 2) and the

School Board's External Research Review Committee (ERRC) and received the necessary ethics approvals from both institutions (see Appendices M & N). I hereby address the ethical obligations involved in this study under two categories: concerns about the data (performance-outcome-related issues) and concerns about authority (consent-related issues).

Concerns regarding the integrity and authenticity of the data are valid in any research project. In this thesis, the students' task performance outcomes on curriculum-based classroom tests constituted a portion of the data that I used. This would have been a concern had my research been focused on using calibration to improve students' grades and if the success of my study depended on the proof of improved grades. Improved academic test performance, although often a by-product of improved metacognition, was not the intention of this study. I studied students' metacognitive calibration, regardless of the decrease or increase in their performance outcomes. Hence, the fluctuations in the students' grades had no bearing on the direction or results of my study. In other words, it was possible to consider a student well-calibrated if he or she performed poorly in the course but could accurately estimate his or her own performance.

The concerns regarding the issues of authority arise when participants are recruited by individuals in a position of authority. In this study, the students' pre-existing entitlement to taking the course was not prejudiced by their decision whether to participate in or to withdraw from this study. Each semester, several sections of the course were offered in the host school where the study was conducted. The students could choose either the section with a research component or another section of the same course without a research component. None of the students left my section of the course with a research component, and one student from a different section requested to switch over to my class. The research procedures were supported by the District School Board and the host school administration. For students who were in the research group, participation in the research was on a voluntary basis; they could stay in the course with or without participating in the research study. Based on the research design, all the consent and assent forms were delivered to the school's main office and I was unaware who the research participants were until the semester ended. The students were informed that they could withdraw from the study any time without any ramifications.

To be effective, metacognitive strategies need to be explicitly taught, modelled, and reinforced by the classroom teacher while teaching (Nietfeld et al., 2005, 2006a), which is a difficult task to achieve unless the research is carried out by the classroom teacher. It is also important to note that this thesis describes a pedagogical intervention designed to be taken up potentially by other classroom teachers. As such, the intervention processes were designed to blend naturally with students' daily classroom routines to become a part of a teacher's day-to-day practice. With these considerations in mind, as Roeder and Roebers (2013) recommended, I opted for authenticity by conducting research with the students whom I taught who might respond more naturally to me as their teacher and to the familiar and relevant classroom material than to an outside researcher who might use unfamiliar research material.

Potential Benefits of This Intervention for Students

In high school, students are often required to learn copious quantities of new material quickly and effectively. Nietfeld et al. (2005, 2006a) explained that the efficacy of students' study behaviour depends, among other things, on the accuracy of their monitoring judgements, because it is based on these judgments that students decide how well a new topic has been understood and when it is appropriate to move on to the next topic. As Nietfeld et al. (2005, 2006a) have detailed, metacognitive strategies do not always develop automatically to a sufficient level and need to be explicitly taught and frequently practiced in order to be effectively internalised and eventually become habitual by students. Throughout the course, I taught the intended metacognitive strategies explicitly and modelled these strategies (see Appendix I for script). The monitoring strategies presented to the students were then mediated through continued conversations, feedback, and personal reflection to build students' metacognitive knowledge towards enhanced monitoring of task performance.

In this era of enforced competitiveness and emphasis on students' academic success, considering the distractions imposed on young people by an increasingly fast-paced lifestyle, social media, and modern-day technologies, teaching high school students some metacognitive strategies becomes an almost ethical issue when research (e.g., Shapiro, 1984; Schunk, 1989; Zimmerman, 1989) demonstrates that by simply asking students to set a goal (e.g., "I am going to learn this chapter very well") and recording an expected result (e.g., "I expect myself to be about 85% accurate on my quiz"), students have displayed superior learning and performance gains. In responding to this ethical calling, I embedded my research methods within the daily classroom activities and invited all the students to participate so they could all benefit from learning metacognitive strategies.

The classroom intervention described in my study might have benefitted the students in many other ways as well. Metacognitive skills, including calibration, are considered foundational to mathematical performance (Desoete, 2009; Kaune, 2006; Lucangeli et al., 1998), particularly in problem-solving scenarios (Allardice & Ginsberg, 1983; Schoenfeld 1982, 1981, 1987b, 1992). The students in the course participated in many episodes of monitoring strategy training. They were encouraged to think carefully and choose the questions that they thought they knew how to answer as their test questions. Students were also engaged in reflection and received feedback. Such activities might have supported their several thinking and problem-solving skills.

General metacognitive strategies transcend grade level and subject area and support students' lifelong learning (Butler & Winne, 1995; Schunk, 1984; Zimmerman, 1983, 2000, 2002). I often communicated this important issue with the students, reminding them that general metacognitive strategies are applicable to both academic and non-academic areas throughout our personal and professional lives. Although I did not collect any data on the benefits of enhanced monitoring processes for students, based on similar studies of metacognition, my teaching experience, and what I have learned from the research literature, I consider that learning metacognitive strategies is critically beneficial for students. In particular, Swanson (1990) emphasised that monitoring constrains cognition by compensating for performance deficits, thereby supporting students' learning and motivation.

Chapter Three Summary

In Chapter Three, I outlined my research methodology and design. To start the chapter, I explained the mixed methods methodology that I employed in this study to assist in providing an answer for my research question. Next, I introduced the research participants, and described the setting, context, and the classroom environment for my study. Further, I outlined the problems associated with the assessment of calibration and detailed the data collection processes and the methods and instruments that I used to collect qualitative and quantitative data. Moreover, I explained the four components of my intervention and addressed the issues related to research quality in mixed methods methodology, discussed the assumptions, limitations, and delimitations of my research design, and highlighted some overall ethical considerations. I concluded Chapter Three by presenting my views on the benefits of my intervention for students and provided a chapter summary. In Chapter Four, I detail my data analysis procedures and the research results.

Chapter Four

Data Analysis and Results

Introduction

In this chapter, I describe the data analysis procedures, detail the findings of my analysis, and provide a summary of the chapter. In mixed methods studies, the data analysis can be particularly challenging because the researcher must engage with both qualitative and quantitative methods of data analysis. Due to the difficulties inherent in mixed methods data analysis, some researchers (e.g., Onwuegbuzie & Combs, 2010) have proposed an inclusive framework that allows the researcher to not only incorporate the data, but also to use the language, terminology, conventions, and analytical methods of both qualitative and quantitative methodologies to answer the research question.

To follow the contemporary trend in calibration research, in my study of students' calibration, I utilised quantitative methods of data analysis including histograms, mean estimates, boxplots, and linear regression models to capture *how* the numerical measures of calibration changed over time. Moreover, to explore and describe *why* the measures of calibration changed as they did, I employed three rounds of blended data (i.e., combined qualitative and quantitative data) analyses—which, to my knowledge, has not been used in prior calibration research. To explain, I started the first round of my blended data analysis by analysing each students' overlapping qualitative and quantitative data. In so doing, I analysed cross sections of the data at the points in time where both qualitative and quantitative data were available for all the students, one student at a time.

For the second round of the blended data analysis, I used the overall blended data for all the students. At this stage, I chronologically compiled both qualitative and quantitative data of the students onto their calibration graphs and then analysed the entire dataset. In these two rounds of data analyses, I looked for emerging ideas about how students described the change in their calibration individually, in the first round, or as a group in the second round of analysis.

In the third round of my blended data analysis, I incorporated the students' voices by adding students' interview data, looking for convergence and consistency in the results, as well as other matters not accessed through using quantitative methods. This triple analysis was coherent with my mixed methods design, in which both qualitative and quantitative analytical techniques were necessary to provide a response for my research question.

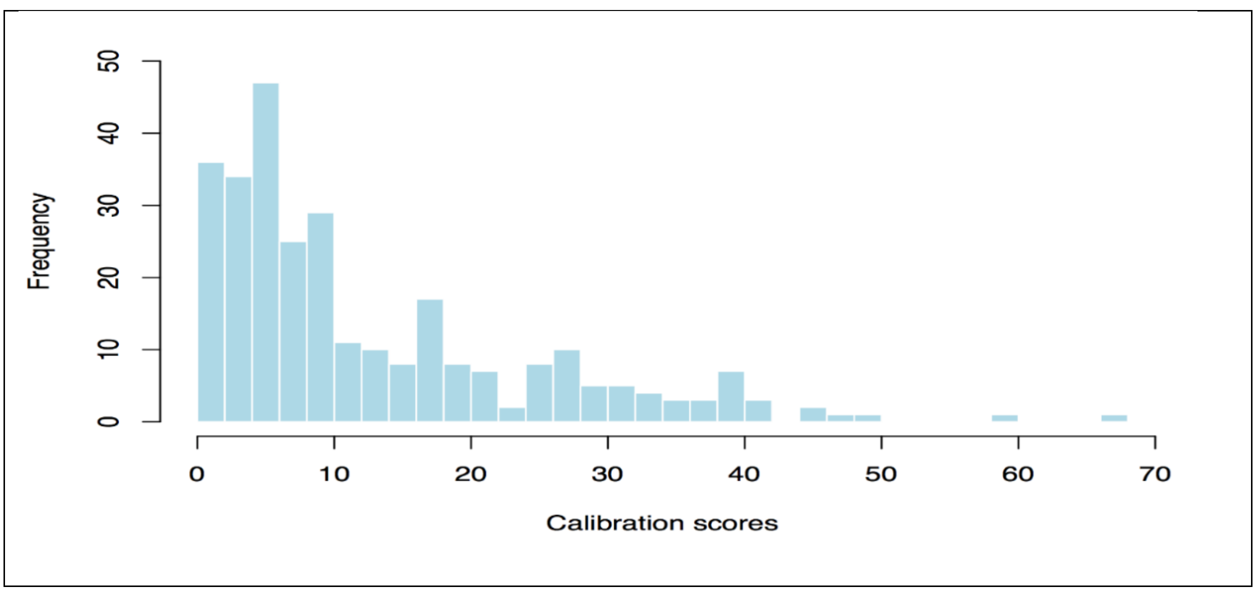
I had not encountered this method of blended data analysis in my review of calibration or mixed methods research studies. However, after trying to analyse my data in different ways and contemplating on what insights those analyses could offer or reveal, I came to understand that a blended data analysis might be more helpful than other methods of data analysis in extracting fresh and constructive ideas from my data. To explain, I conceived this novel way of data analysis after I realized that my data possessed a particular property that I had not previously encountered in my review of calibration studies: the simultaneous availability of the qualitative and quantitative data at certain times (i.e., the testing episodes). Moreover, my appreciation for the social and developmental aspects of calibration made me wonder whether the factors of time and social interactions were more important to the notion of change and development of students' calibration than I had already envisioned. Hence, I decided to merge and analyse the blended qualitative and quantitative data instead of separately analysing each data set. I begin Chapter Four by describing my quantitative data analysis procedures.

Quantitative Data Analysis

In this study, the importance of quantitative analysis lies in the centrality of the notion of change in students' calibration to my research question. I started my quantitative analysis by exploring how the numerical values of students' calibration changed over time and across 12 tests, and what patterns emerged from the quantitative data. The variable of interest was calibration. Figure 4.1 represents the combined distribution of the measures of calibration for 12 tests for all students. Since calibration values (i.e., called calibration scores in Figure 4.1) mark the gap between estimated and achieved performance outcomes, the lower the calibration values, the closer the estimated and achieved performance outcomes, which indicates better calibration. In Figure 4.1, the heavy skew on the left side of the histogram where the measures of calibration are low is indicative of the overall prevalence of good calibration for 24 students over 12 tests (*n* = 288). For instance, the histogram shows that the calibration scores were as low as 0% – 2% in 32 out of 288 measured calibration scores.

Figure 4.1

Histogram Representing the Distribution of the Combined Calibration Scores for all Students Over 12 Tests

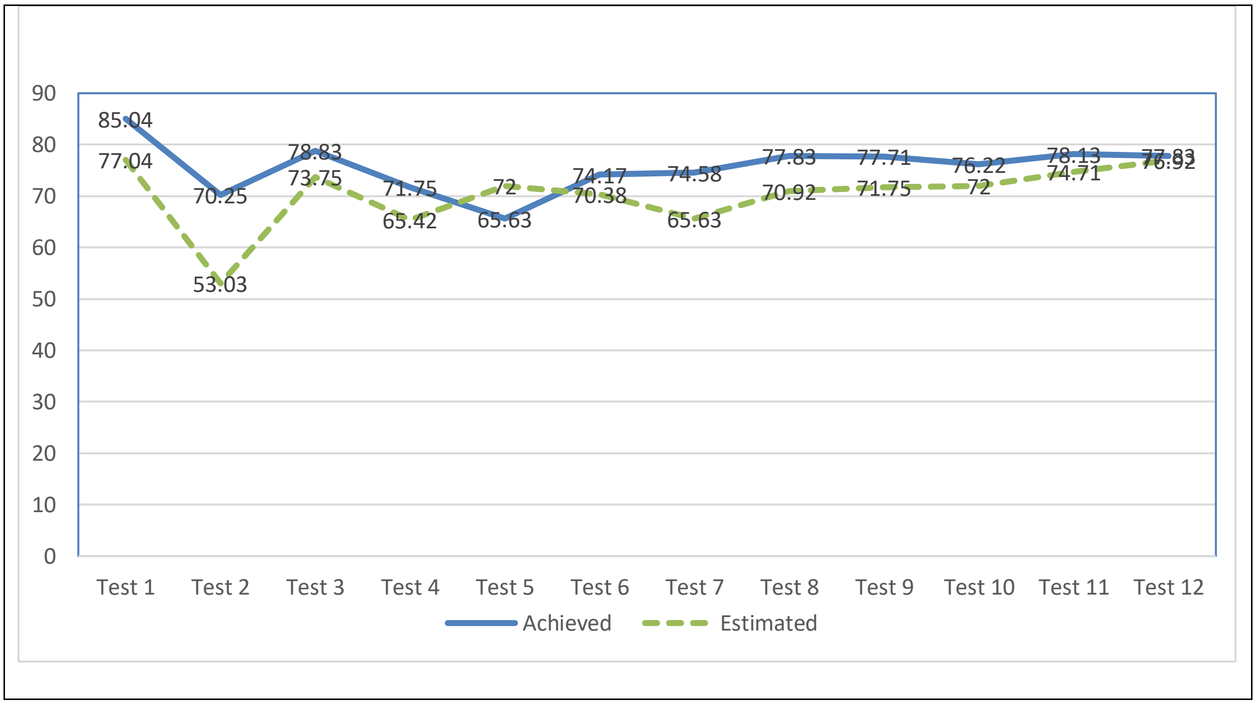


Further, in intervention studies within calibration research with a large number of tests, it is not unusual to compare the results of some of these tests. For instance, Brookhart et al. (2004) compared the results of the last five weekly quizzes out of a total of 10 quizzes, and Roderer and Roebers (2013) compared the results of the fifth, eighth, and tenth out of 10 tests as a measure of how effective their respective intervention had been. In this study, to examine the potential effects of my intervention on students' calibration and to explore students' calibration change, I compared the students' calibration accuracy at the beginning using the first test, to the students' calibration accuracy after the intervention using the last test. Since the same participants took part in all classroom tests and the dependent variable (i.e., calibration accuracy) was the same across the tests, I applied dependent *t*-tests to the calibration scores on Test #1 at the beginning of the intervention versus Test #12 in the end, using the IBM SPSS 24 software. Theoretically, a *t*-test determines whether there is a significant difference between the measures of two groups. For this data set, I calculated the Cohen's effect size d (Cohen, 1988, 1992) as d = .815, which indicates a large effect. To explore this line of inquiry further, I applied a regression model, comparing a subset of the first two tests (Test #1 and Test #2, $R^2 = .5063$) relative to the last two tests (Test #11 and Test #12, $R^2 = 0.8071$). This analysis revealed a substantial contrast of 30% from the beginning of the semester relative to the end, as the predictability of calibration accuracy by using multiple testing episodes rose from 50.63% to 80.71%, , which is consistent with similar findings (e.g., Roderer & Roebers, 2013) in calibration research.

To explore how calibration changed for the whole group throughout the study, I compared the students' mean performance with their mean estimation of performance. Figure 4.2 represents the combined mean performance percentages versus the mean estimations of performance, also as a percentage, for all students. As Figure 4.2 illustrates, as a group, the students began by underestimating their performance and gradually became more accurate in their self-assessments of task performance.

Figure 4.2

All Students' Combined Mean-Estimate Versus the Mean-Achieved Test Performance Scores Over 12 Tests



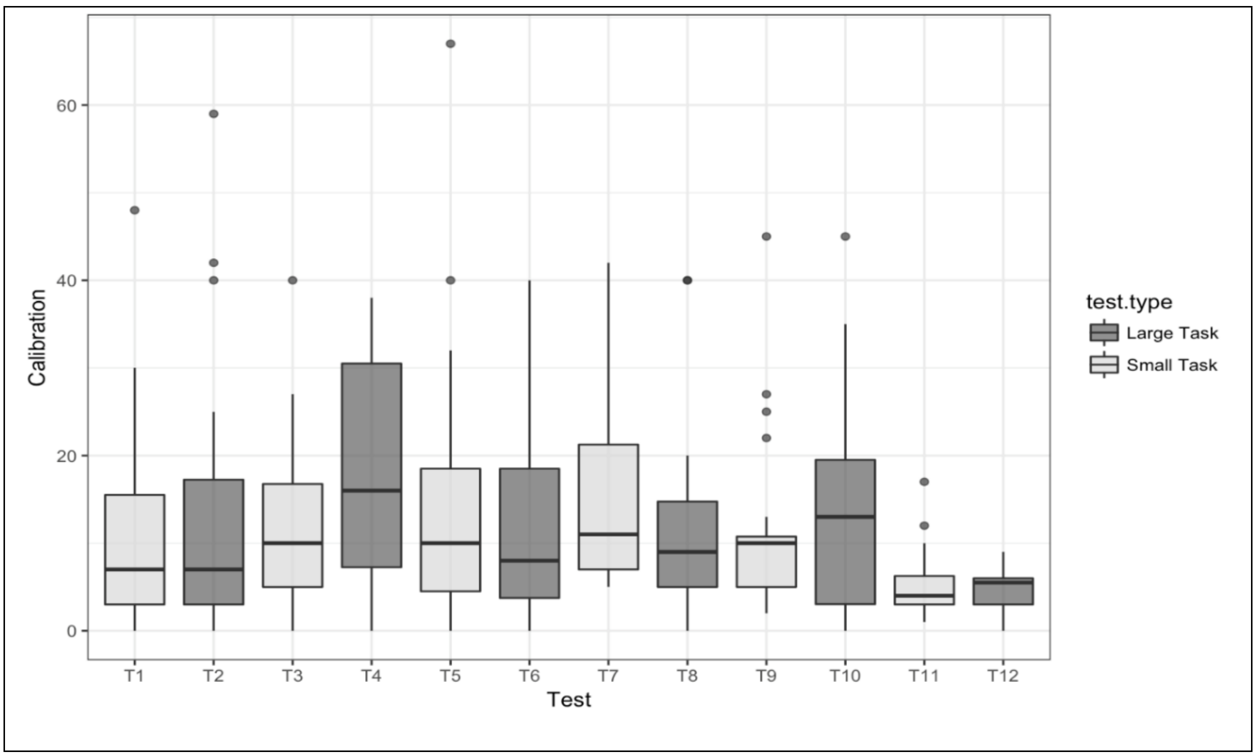
To test for the systematic differences between estimated and achieved test scores, I performed a series of 12 pairwise dependent *t*-tests. These analyses are generally recommended (e.g., Field, 2009) for small sample population size and when the same individuals participate in research across a variety of testing conditions. Field (2009) states that when the same participants take part in research across different conditions, the unsystematic variance is reduced dramatically, allowing for a much easier detection of systematic variance. The results of the pairwise dependent *t*-tests suggested that only for Test #1 and Test #11, the mean test performance estimations were substantially different from the mean test performance scores ($t_{test1}(24) = 2.273$, $p = .033$; and $t_{test\ 11}(24) = 2.092$, $p < .05$). For all the other tests, the differences between the means were not significant ($.07 < p < .87$). This was important for me to investigate because calibration is inherently related to task performance. Since the focus of this study was not students' performance scores, I cannot relate the overall improved performance scores to the change in students' calibration. However, the results were also reassuring in terms of the stability of students' performance outcomes.

Except for the outliers, the measures of students' calibration accuracy ranged from absolute value of 0% to approximately 40%. The median calibration score was equal to 9%, indicating that 50% of students demonstrated an overall calibration measure of less than 9%. The standard deviation of the measures of calibration decreased substantially towards the end of the semester, in particular for Test #11 and Test #12. This outcome indicated that in the context of the applied intervention, the variability in calibration scores decreased. In other words, the gap between the most-calibrated students and the least-calibrated students narrowed—a welcome result, considering that the overall calibration accuracy of students also improved over time.

As noted earlier, all the tests were written within the four achievement categories of Knowledge & Understanding, Application, Communication, and Thinking. The weights of these categories were 20%, 20%, 15%, and 15% respectively and the culminating project and the final exam had a combined weight of 30% toward the overall term mark. I explored whether the type of the test had any effect on the students' calibration accuracy. In so doing, I used a boxplot to examine the distribution of students' calibration measures per test on small tasks (quizzes with 6 questions) and on large tasks (unit tests, the midterm exam and the final exam with 10, 15, and 30 questions, respectively), as Figure 4.3 depicts. Field (2009) describes boxplots as five-number summaries that are especially useful when comparing two or more data sets. In a boxplot diagram, the two ends of each box represent the upper and lower quartiles. The horizontal line inside the box represents the median and the lines extended outside the box represent the highest and lowest data points. My analysis suggested that except for the outliers, the task type had no significant effect on calibration accuracy that was worthy of further investigation.

Figure 4.3

Boxplot Representing the Distribution of Calibration Measures Categorised by Task Over 12 Tests



In statistical studies of calibration, in determining whether there are significant error variances, one of the most commonly reported factors is the Wilks's lambda. Reported on a scale of 0 to 1, the value of this statistic determines the ratio of error variance to total variance for each variate. A Wilks's lambda of 0 is ideal (Field, 2009) because it means that there is no variance which cannot be explained by the variable under examination. In this study, for the overall test of significance, I computed the Wilks's lambda as .272 (see the first value on the second line of Table 4.1), which, based on Cohen's (1988) guidelines, reflects a large effect size. Further, a (partial) eta^2 value of .728 (see the sixth value on the second line of Table 4.1) suggested that 72.8% of the variance in the measures of the students' calibration accuracy could be explained by the 12 tests that the students took. In other words, this analysis suggested that 72.8% of the overall improvement in students' calibration was related to the classroom intervention that took place.

Table 4.1

Wilks's Lambda for Overall Test of Significance

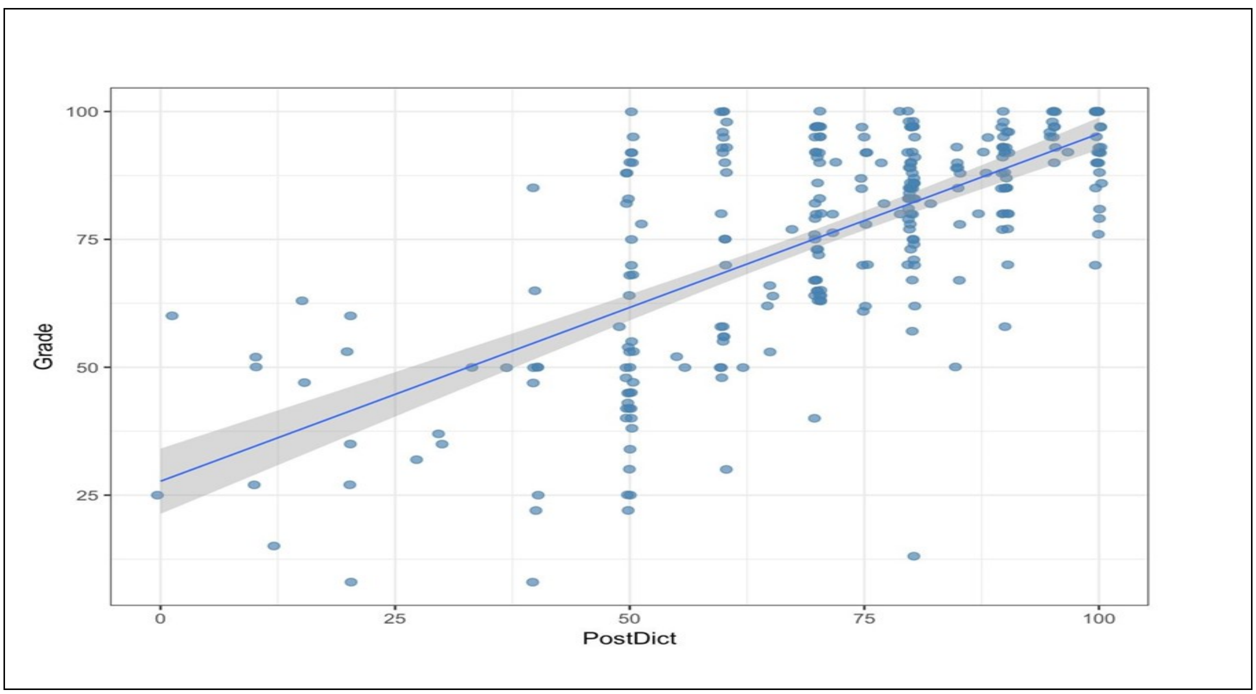
	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Observed Power
Pillai's trace	.728	2.925	11.000	12.000	.039	.728	.788
Wilks's lambda	.272	2.925	11.000	12.000	.039	.728	.788
Hotelling trace	2.681	2.925	11.000	12.000	.039	.728	.788
Roy's largest root	2.681	2.925	11.000	12.000	.039	.728	.788

Each F tests the multivariate effect of calibration change. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

The simple linear regression model of the performance outcomes versus performance estimations resulted in $R^2 = .4543$, with df = 286, F = 238.1, and $p < 0$. The value of R^2 suggests that 45% of the variance in the students' performance outcomes may be predictable from the estimations on 12 tests that they took. The relatively low R^2 (low predictability) might be explained by some students' miscalibration. It is possible that had the students not given themselves several low performance estimations, the discrepancy between the estimated and achieved could have been reduced, possibly resulting in a higher measure of predictability. Task type (quizzes versus unit tests and exams) had no effect on the model and was removed as an independent factor for a better fit. Figure 4.4 depicts this linear regression model with the standard error of the mean (with 95% confidence interval) visualised in grey surrounding the regression line.

Figure 4.4

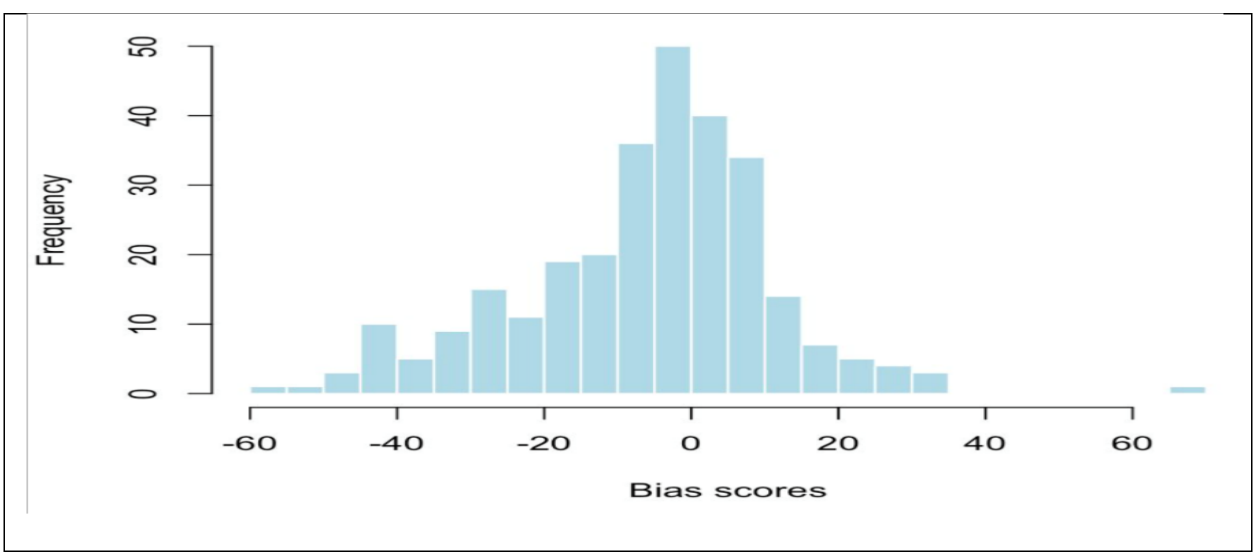
Scatterplot of the Linear Regression Model for Estimated Versus Actual Performance



As I noted in Chapter Three, calibration is numerically evaluated as an absolute value; as such, it is always reported as a positive number. As a classroom teacher, I found it informative and helpful to explore students' directional miscalibration (i.e., bias) and investigate whether students were overestimating (i.e., positive bias) or underestimating (i.e., negative bias) their task performance. I employed a simple linear regression to examine the change in the measures of bias over 12 tests. The regression line was generated as y = -7.13 + 0.33x, and the model summary provided $R^2 = .196$, suggesting that almost 20% of the intended result (i.e., a decrease in bias) was accounted for by the repeated testing sessions. This issue suggests that the development of students' calibration through classroom interventions may require time and several training sessions to take effect. Further, this may explain why conducting one-time, snapshot style research on calibration may produce inconclusive or contradictory results. With respect to the significance of this correlation for the dependent variable (i.e., measures of bias), the coefficient table on SPSS 24 provided $p = .001$, and the lower and upper bounds for bias were calculated as – 10.580 and – 3.677, respectively, with a confidence level of 95%. Figure 4.5 represents the distribution of the measures of bias ($n = 288$), illustrating that the students predominantly demonstrated negative bias by underestimating their task performance outcomes.

Figure 4.5

Histogram Representing the Distributions of Bias Measures



To further investigate students' directional miscalibration (i.e., bias) and to highlight the range of bias scores as well as the mean values and the standard deviation for the obtained values of bias, Table 4.2 provides the descriptive statistics of the bias scores across 12 tests demonstrating the trend of decreasing variation (i.e., smaller values of standard deviation) over time. Table 4.2 illustrates that as a group, students consistently underestimated their task performance outcomes on mathematical tasks and this underestimation gradually became less pronounced over the 12 testing episodes.

Table 4.2

Measures of Bias for All Students Over 12 Tests (n = 288)

	N	Minimum	Maximum	Mean	Standard Deviation
T1_Bias	24	-48	19	-8.25	16.329
T2_Bias	24	-59	25	-6.83	19.395
T3_Bias	24	-40	25	-5.08	16.365
T4_Bias	24	-38	30	-5.92	21.069
T5_Bias	24	-40	67	-0.50	21.456
T6_Bias	24	-42	32	-3.25	19.802
T7_Bias	24	-50	30	-8.96	18.454
T8_Bias	24	-45	28	-6.92	18.406
T9_Bias	24	-45	30	-5.96	16.470
T10_Bias	24	-45	35	-3.96	18.314
T11_Bias	24	-25	10	-3.42	8.000
T12_Bias	24	-9	9	-0.92	4.492
Valid N (listwise)	24				

In quantitative analysis, a parameter called the *residual statistic* is often used to test the assumption of *homoscedasticity*. Field (2009) explains that the measure of residual statistic speaks to the soundness of a regression model. The test of homoscedasticity, which was particularly important to the present study due to the small number of participants, determines whether the residuals of a regression analysis have the same variance (which indicates homoscedasticity), or they have changing variance (which indicates heteroscedasticity). In analysis of the regression residual, it is hoped that the variance of the residual terms remains constant (Field, 2009), resulting in the mean residual to be zero. Table 4.3 shows that the residual statistic for the regression model applied in this study with the mean residual equal to zero.

Table 4.3

Residual Statistics for the Measures of Bias

	Minimum	Maximum	Mean	Std. Deviation	N
Predicated Value	-6.7999	-3.1868	-4.9933	1.18428	12
Residual	-4.13090	4.98603	.00000	2.39962	12
Std. Predicted Value	-1.525	1.525	.000	1.000	12
Std. Residue	-1.641	1.981	.000	0.953	12

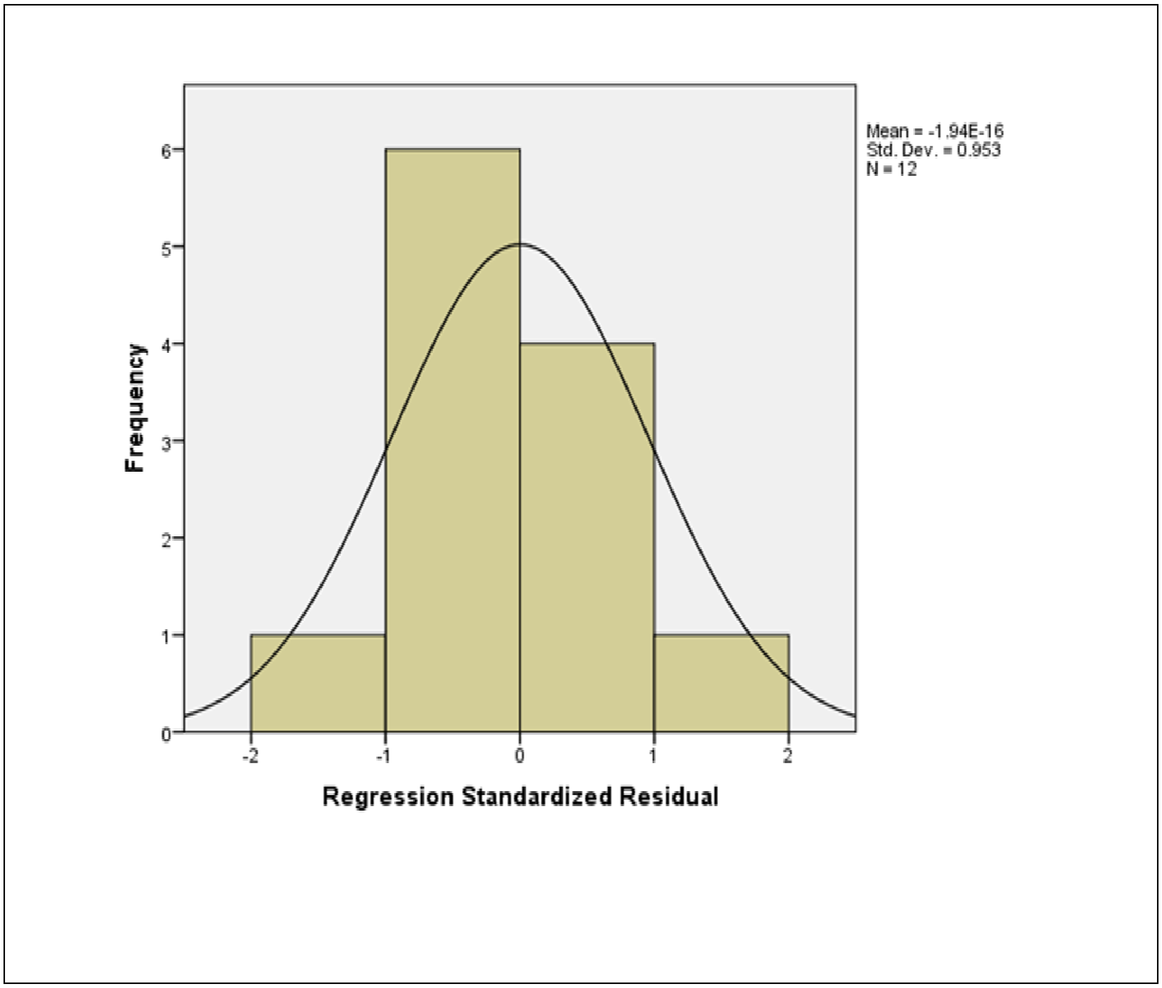
a. Dependent Variable: Mean Bias

While the model that tested students' performance estimations (i.e., postdictions) was not homoscedastic in a sense that the variance around the regression line was not the same everywhere (see Figure 4.4), the model testing bias was homoscedastic (see Table 4.3). The homoscedasticity of the model testing bias suggests soundness of the applied regression model and indicates better predictability than heteroscedastic models.

Field (2009) proposes that for the regression analysis to be a good fit, it is required for the regression standardised residual distribution to be approximately normal. As such, the residual distribution for the regression analysis used in this study was tested and the condition of normality was met indicating that the regression model was a good fit. Figure 4.6 illustrates the residual distribution within the regression model used in this study.

Figure 4.6

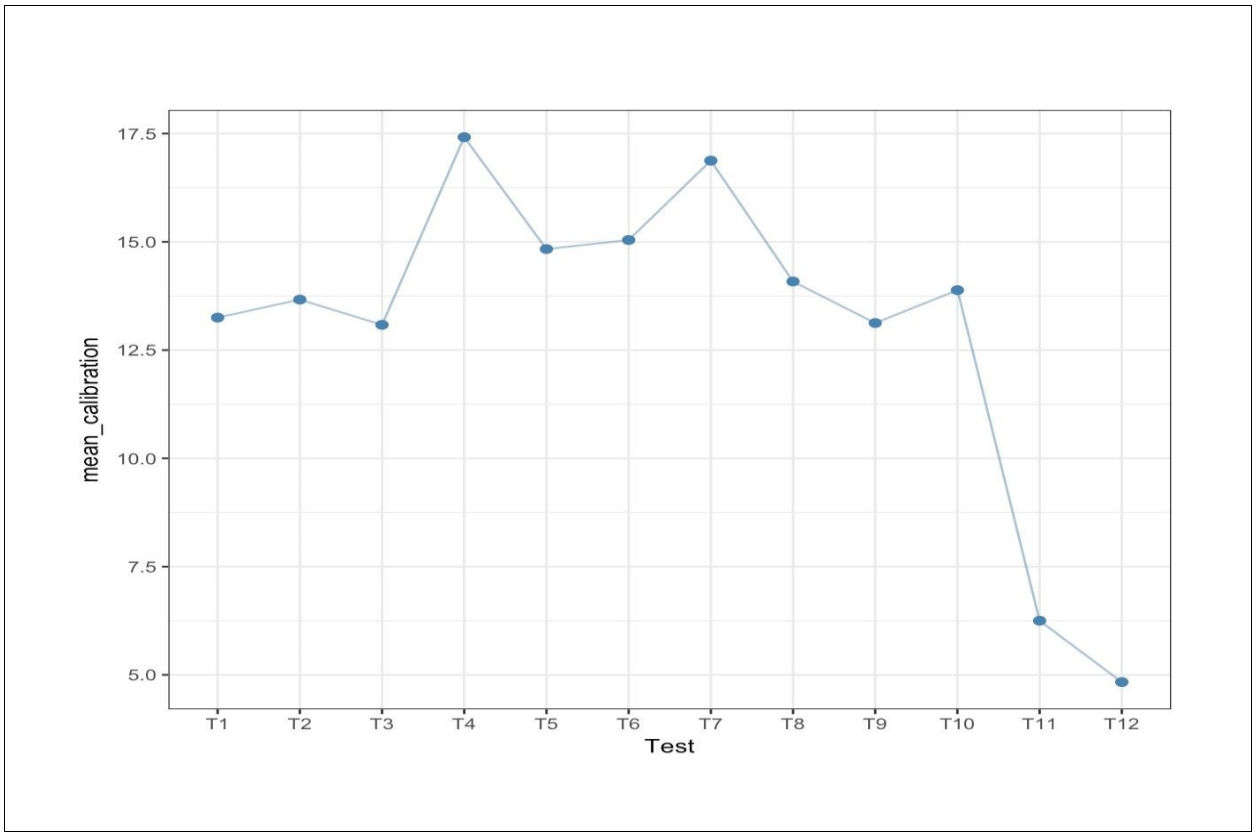
The Residual Distribution Within the Regression Model



The quantitative analysis of my data provided me with the opportunity to explore the notion of calibration change, which was central to my study. Figure 4.7 represents the change in the mean calibration scores over 12 classroom tests. As illustrated in Figure 4.7, there is a notable decrease in the measures of calibration (i.e., calibration scores) after Test #7, toward the end of the semester. Since the measures of calibration indicate the gap between students' estimated and achieved performance outcomes, decrease in calibration measures toward the end of the semester suggests that over time students' calibration accuracy gradually improved.

Figure 4.7

Scatterplot of Mean Calibration Scores Over 12 Tests

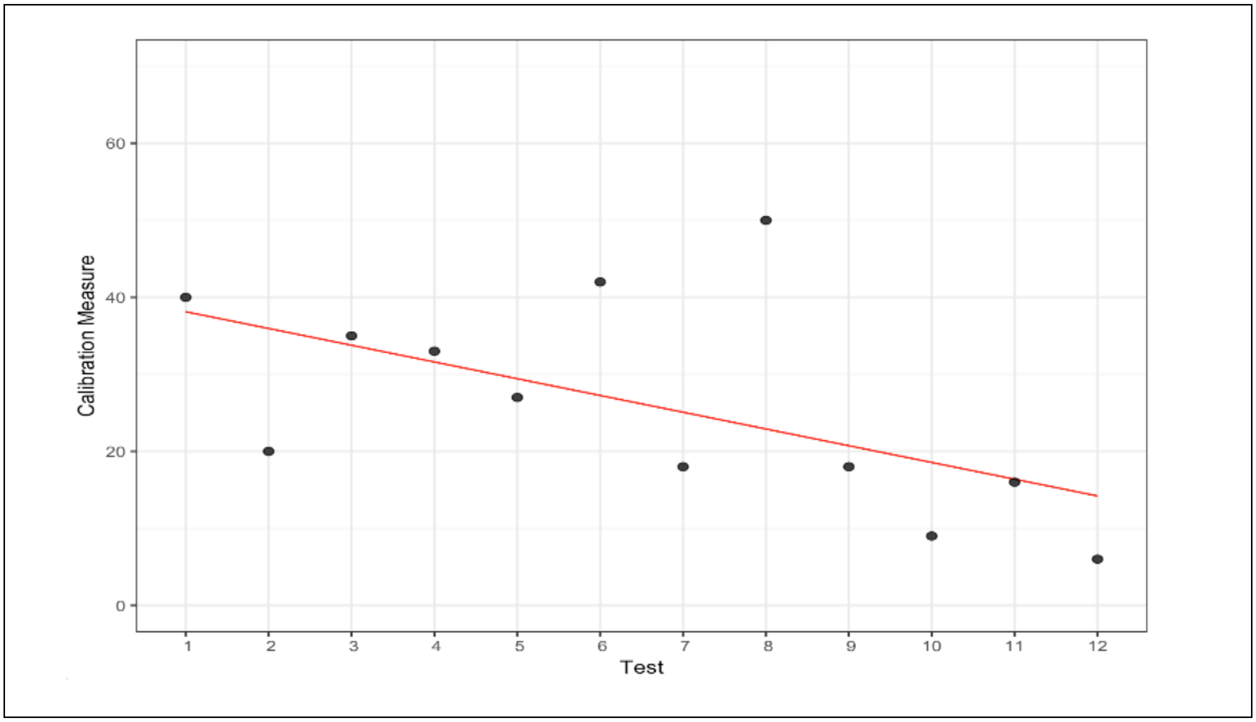


The quantitative analysis, as outlined above, was an investigation of the patterns of change in students' overall calibration as a group. Exploring the change in individual student's calibration, however, was immensely important to this research, conceptually, philosophically, and practically. Researchers (e.g., Gutierrez et al., 2016) posit that generally people are known to possess some level of monitoring skills. Further, calibration researchers (e.g., Pieschl, 2009) have suggested that monitoring judgments which are foundational to an individual's calibration accuracy may reflect some aspects of an individual's stable personal characteristics that have little to do with monitoring. Zhao and Linderholm (2008) suggest that in making monitoring judgments, individuals might first anchor their judgment with their (often unrelated) past experiences and then adjust their monitoring judgment based on the characteristics of the actual task at hand. Rutherford (2017b) explains that despite the adjustments made, the monitoring judgments are still biased toward individuals' more stable personal characteristics. Hence, it was important for me to employ a mixed methods approach and explore the calibration accuracy of individual students over several tasks to distinguish for instance between task-specific monitoring judgments (e.g., "I have not practiced with enough equation-solving questions, so I should estimate low"), and the individual's stable personal characteristics (e.g., "I am not a math person. My test results will always be low, so I should estimate low").

Appendix K represents the change in individual student's calibration over time. In each graph, the y-axis represents each individual student's calibration measures, while the x-axis represents the 12 tests the student took. Further, the line of best fit depicts how individual student's calibration accuracy changed across 12 tests. Since calibration is a positive number, a descending line of best fit refers to an improvement in individual student' calibration accuracy as the numerical values of calibration decrease over time, whereas a horizontally-situated line of best fit depicts a lack of change in the calibration measures over time. An ascending line of best fit illustrates an increase in the numerical values of calibration measures over time, indicative of a decline in calibration accuracy. The size of these 24 individual calibration graphs in Appendix K has been reduced to save space. To provide an example, Figure 4.8 depicts Adnan's calibration graph. In this graph, the numerical values of Adnan's calibration are represented on the y-axis and the 12 classroom tests are depicted on the x-axis. The descending line of best fit indicates that Adnan's calibration accuracy improved over time.

Figure 4.8

Adnan's Calibration Graph Over 12 Testing Sessions



These graphs depict that the numerical measures of calibration decreased for 18 out of 24 students. Moreover, Lusine, Sun Xin, Marel, and Wang with initially good calibration measures at the start of the semester remained highly calibrated with a calibration measure within 5% throughout the semester. These students represent examples of the ceiling effect in a sense that since they were initially well-calibrated, there was not much room for improvement for them. My quantitative analysis suggested that of 24 students, 22 students possibly benefitted from the intervention and either became better calibrated or stayed at a good level of calibration accuracy over time, across the 12 testing episodes. However, two students, Rabia and Bilal became less calibrated over time. My analysis suggests that as a group, the gap between the most calibrated students and the least calibrated students narrowed over time.

Blended Data Analysis

In this study, I took the less-travelled path of employing both qualitative and quantitative methods to explore the ways in which the calibration of a group of students changed over time. My quantitative analysis which provided me with insights into the malleability, complexity, and variability of calibration was an important aspect of my study. To explain, calibration is often referred to as the numerical disparity between the estimated and achieved performance outcomes (e.g., Bol & Hacker, 2012), which can be explored only quantitatively. Although as a cognitive social phenomenon, the popularity of calibration as a research topic is on the rise (Rutherford, 2017a; Stoten, 2019), calibration research findings are mainly based on a predominantly quantitative approach to research.

My quantitative analysis highlighted for me not just the change in calibration, but also the dispersion and variability in the numerical measures of calibration over time. For me, however, questions remained regarding the nature of calibration change, and the reasons for its complexity and variability. For instance, although the quantitative data showed that overall, as a group, students' calibration accuracy improved over time, I wondered:

- What factors might have influenced students' calibration?
- Were these factors intrinsically or extrinsically oriented?
- How did the students make sense of calibration and what it meant?
- Was the students' sense of well-being affected in any way by my research?
- What were the bases of the students' self-estimations?
- What were the students' views about their experiences with this intervention?

- What were some of the reasons for students' miscalibration from their own perspective?
- What were some of the reasons for students' miscalibration based on the data?

As a researcher and classroom teacher, I was intrigued by what I could do to seek answers to these non-quantitatively oriented questions. I also realized the unique quality that my data possessed in a sense that after making performance estimations (i.e., after a test), students also wrote in their reflection notecards providing me with 12 sets of simultaneously generated quantitative and qualitative data. To take advantage of this unique situation, working in electronic files, I chunked my qualitative and quantitative data on a timeline that represented the progress of my classroom intervention. I then started analysing the blended data chronologically, once for each individual student, and a second time for the group.

Overall, I performed three rounds of blended data analysis. In the first round, I started my analysis by using 12 time-stamped instances which I termed *cross sections* of the blended qualitative and quantitative data for each student. These were the 12 testing episodes for which I had collected both qualitative and quantitative data for each student. I found this practice, which I termed the analysis of the *within-student* influences on calibration, essential to my understanding of the change in individual student's calibration. At these 12 cross sections, I examined the combinations of each individual student's qualitative and quantitative data in the chronological order that the student had created the data throughout the intervention.

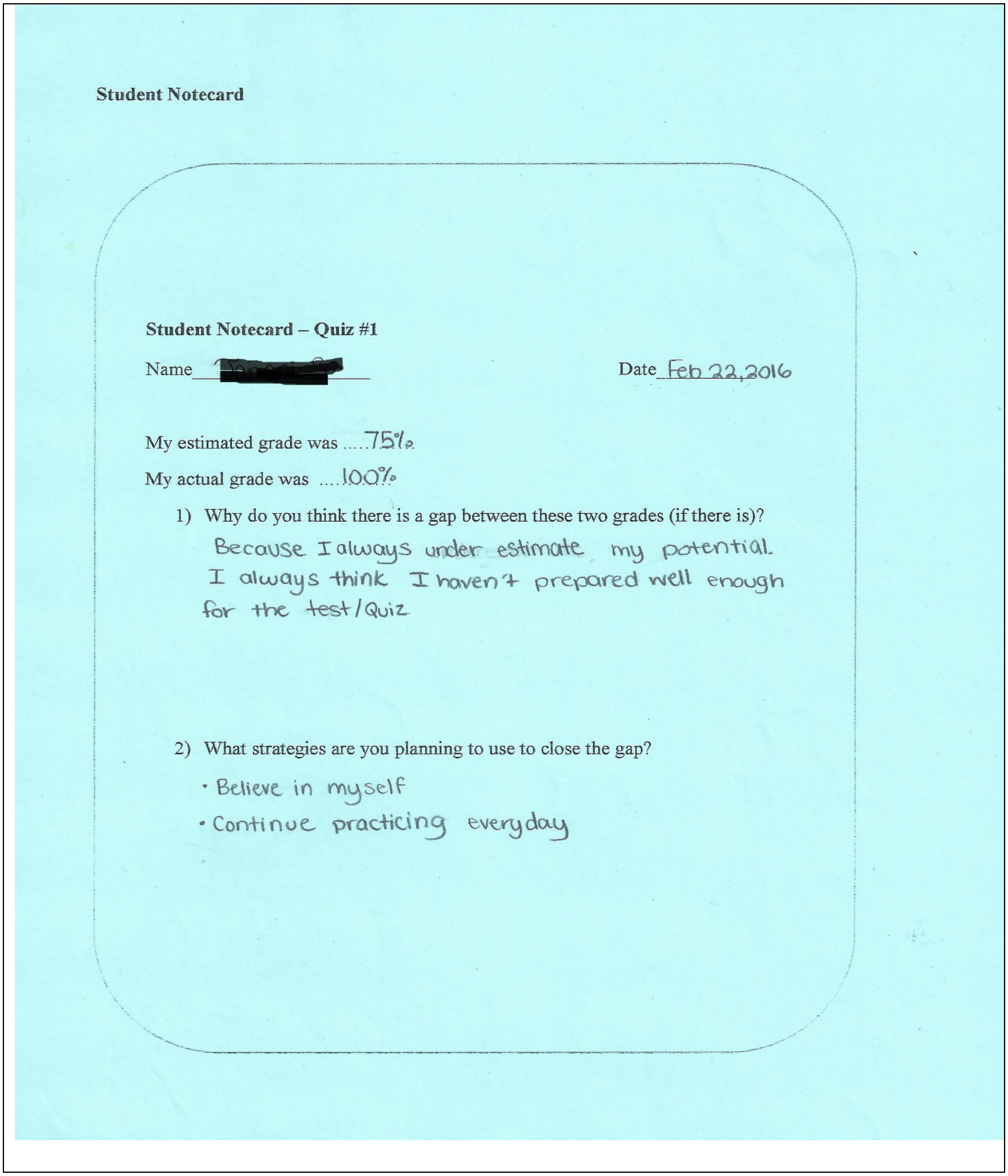
In social cognitive theory (Bandura 1986), forming habits and behaviours (including self-monitoring habits) depends on the reciprocal interactions of person, behaviours, and environments. As such, In the second round of the blended data analysis, I looked for what influenced the students' calibration as a group. In this practice, which I termed the analysis of the *across-students* influences on calibration, I considered the blended qualitative and quantitative data and used the time stamped blended data of all the students throughout the intervention.

As I explained in Chapter Three, I did not use the qualitative interview data for the first two rounds of my blended data analyses. My intention was to save these data to be used as a final layer of analysis. Hence, in the third and final round the blended data analysis, I incorporated the interview data to seek convergence, confirmation, and consistency in students' written and spoken data.

Consistent with Bandura's (1986) social cognitive theory, in my blended data analyses I took into account the quantitative measures of students' calibration as well as students' intimations about the personal, behavioural, and environmental factors which influenced students' calibration according to them. To elaborate, in this study, students completed two open-ended questionnaires, one at the beginning and one in the end of the intervention. Further, throughout the intervention phase and simultaneous with completing their calibration graphs, students reflected on their calibration and completed individual reflection notecards at the end of each of their 12 tests. As noted earlier, on each notecard students responded to two questions: 1) Why do you think there is a gap between your estimated and achieved grades, and 2) What strategies are you planning to use to close the gap? It is important to clarify that during my monitoring strategy training sessions, in teaching about calibration, and in overall classroom conversations, we (students and myself) used the term *gap* in a specific way: to refer to the discrepancy between students' estimated and achieved performance outcome, and the phrase *to close the gap* was used synonymously with the phrase *to become better calibrated*. Figures 4.9 and 4.10 illustrate two samples of students' reflection notecards that I used in analysis.

Figure 4.9

Neha's Reflection Notecard



Student Notecard

Student Notecard – Quiz #1

Name [redacted] Date Feb 22, 2016

My estimated grade was 75%

My actual grade was 100%

1) Why do you think there is a gap between these two grades (if there is)?

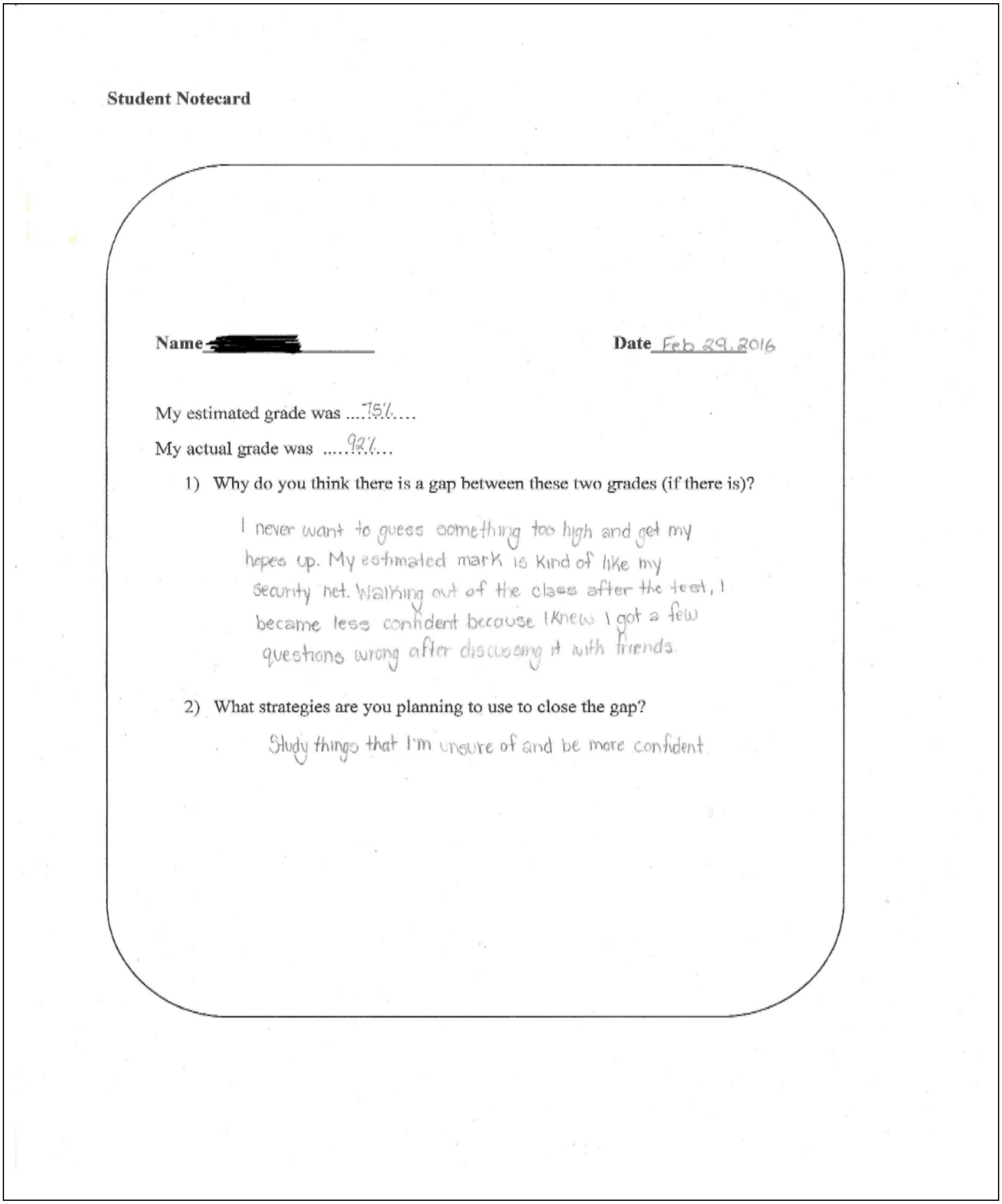
Because I always under estimate my potential. I always think I haven't prepared well enough for the test/Quiz.

2) What strategies are you planning to use to close the gap?

- Believe in myself
- Continue practicing everyday

Figure 4.10

Fang's Reflection Notecard



Student Notecard

Name [redacted]　　　　**Date** Feb 29, 2016

My estimated grade was ....75%....

My actual grade was .....92%....

1) Why do you think there is a gap between these two grades (if there is)?

I never want to guess something too high and get my hopes up. My estimated mark is kind of like my security net. Walking out of the class after the test, I became less confident because I knew I got a few questions wrong after discussing it with friends.

2) What strategies are you planning to use to close the gap?

Study things that I'm unsure of and be more confident

Students' comments on reflection notecards were transposed, one student at a time, onto students' individual calibration graphs which contained their quantitative calibration data. To achieve this, I printed and enlarged students' calibration graphs, which allowed enough space to transpose students' assertions, at the times that they were made, on the calibration graphs, side by side with the quantitative measures of calibration at the time, within the overall pattern of calibration change (i.e., the shape of the calibration graph). The result was a set of blended qualitative and quantitative data, which I used for analysis. Moreover, considerations were made for the Questionnaire #1 data to be added to the beginning of each student's graph and for the Questionnaire #2 data to be added to the end of each student's graph consistent with the timeline in which the data were created by each individual student. I analysed the resultant blended data for personally, behaviourally, and environmentally oriented influences on students' calibration.

Within-Student Blended Data Analysis

To explore the within-student influences on calibration, I proceeded on each student's calibration graph from left to right, chronologically, and considered 12 vertical cross sections of the blended data at the points where both the qualitative and quantitative data were available for each student. These points were, in fact, the testing episodes for which I had the student's quantitative measure of calibration, as well as their qualitative data of reflection notecards. I added the data from Questionnaire #1 to the first cross section at the beginning of the intervention and the data from Questionnaire #2 to the last cross section in the end of the intervention. Subsequently, I moved from left to right on each calibration graph, chronologically exploring how and why, according to students' words, calibration changed for each student, qualitatively and quantitatively, as time passed by.

Examples of my Within-Student Blended Data Analysis. In this section, I present four examples of my within-student blended data analysis for Neha, Anna, Rabia, and Fang in which I analysed 12 cross sections of the blended qualitative and quantitative data for each student. To save space, seven more examples of my within-student blended data analysis are presented in Appendix L.

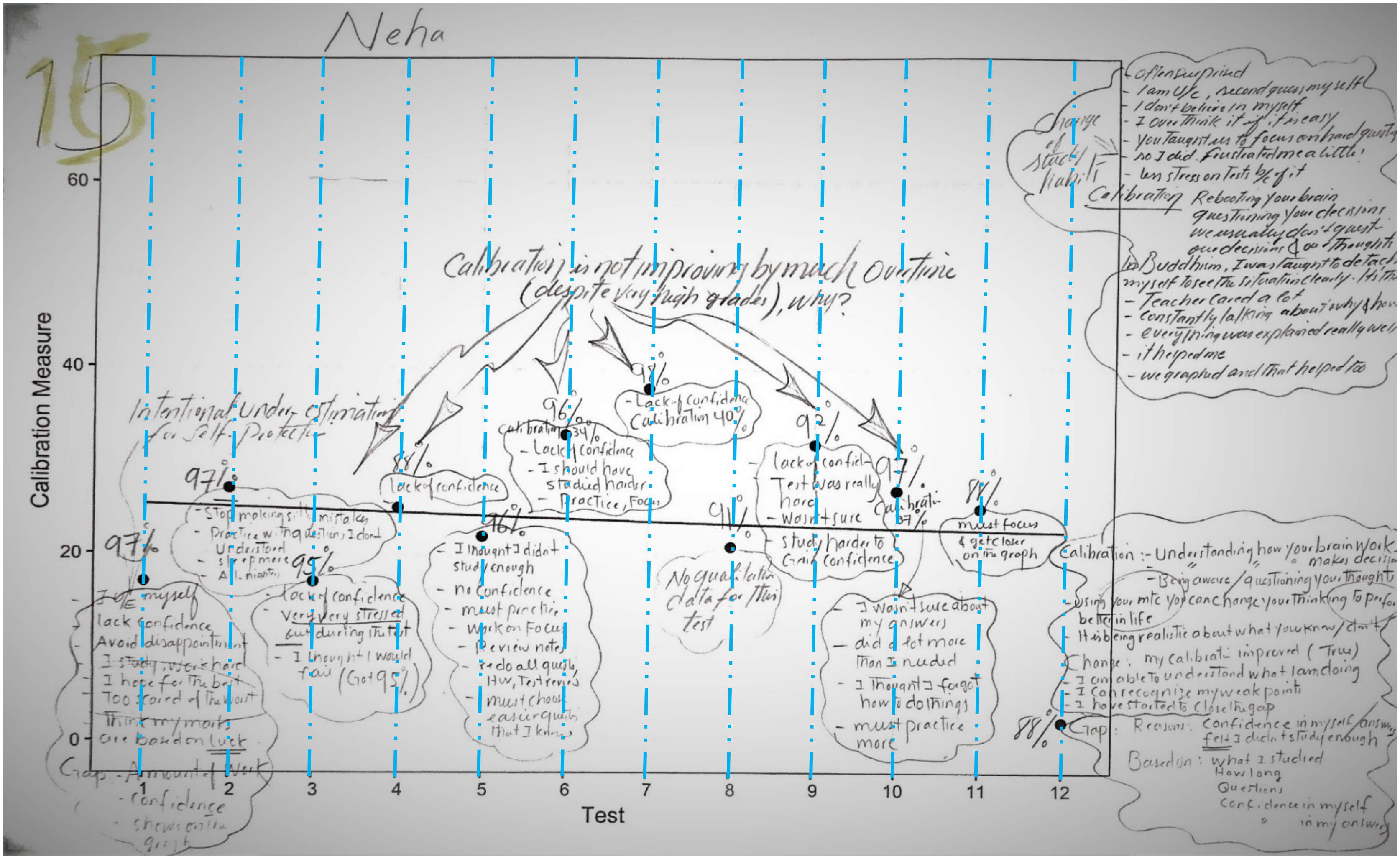
The Example of Neha. As illustrated in Figure 4.11, I merged Neha's qualitative and quantitative data and transposed her blended data onto Neha's calibration graph. For clarity, I organised students' comments in cloud-like shapes to distinguish between students' comments and my own analysis notes.

To explain further, the 12 testing episodes were the times at which I had each student's both qualitative and quantitative data available to me. Hence, for example in the case of Neha, I transposed Neha's written comments regarding her calibration onto her actual measures of calibration at the time of each of her 12 tests. This style of analysis afforded me 12 cross sections of the blended data which included simultaneous qualitative and quantitative data of each student. My intention was to explore, from Neha's perspective, what possibly influenced her performance estimation and hence her calibration at the times that she made her performance estimations.

Hacker and colleagues (2000) propose that monitoring instructions improve the calibration accuracy of high achieving students, such as Neha. However, as illustrated in Figure 4.11, the values of Neha's miscalibration remained high throughout the intervention phase, and overall, Neha's calibration accuracy improved by only 2% which is negligible. Figure 4.11 represents the 12 cross sections of the blended qualitative and quantitative data (illustrated as vertical dashed lines) that I used to analyse the data for Neha.

Figure 4.11

Neha's Blended Qualitative and Quantitative Data Used in Analysis



As viewed in Figure 4.11, Neha achieved high marks on her tests. However, despite her achievements, she remained highly miscalibrated, as the gap between her estimated and achieved grades ranged between a minimum value of 17% and a maximum value of 38%. Calibration literature states that high achieving students are well-calibrated (e.g., Hacker et al., 2008b; Zabrucky, 2010), which presupposes that students like Neha are accurate self-evaluators. My blended data analysis, however, suggested that Neha was not well-calibrated and provided possible explanations as to why this might be so.

As noted in research design, every week on reflection notecards, students were asked about the reason why there might be a gap between their estimated and achieved grades of the week and how they were planning to close the gap. In the context of this study, the notion of gap was used to refer to the discrepancy between students estimated and achieved performance outcomes. In this regard, Neha's comments about closing the gap of miscalibration suggested that Neha's main issue was her self-confidence. For instance, on Notecard #6 with a gap of 36% (estimated 60% and achieved 96%), Neha explained, "[The reason for the gap is] lack of confidence." On Notecard #7 with a gap of 38% (estimated 60% and achieved 98%), Neha gave the same reason, "The reason for the gap is lack of confidence. I was underconfident about my answers and I thought I forgot how to do stuff. I did not know if I was doing it right." Although Neha used the terms 'confidence' and 'self-confidence,' her written comments occasionally referred to what might be considered as her self-efficacy beliefs. For instance, on Notecard #7, when she wrote, "I did not know if I was doing it right," Neha was reporting on her self-efficacy beliefs.

Overall, the cross-sectional analysis of Neha's blended data revealed that despite her high test scores as indicated by the quantitative data, Neha reported about being underconfident both in herself (self-confidence beliefs, such as “I felt underconfident in myself” on Notecard #1), as well as in the outcome of her efforts (self-efficacy beliefs, such as “I was underconfident about my answers” on Notecard #10). Further, Neha experienced stress. For instance, on notecard #4, she commented “I have to sleep more to reduce stress”. Moreover, Neha thought that having the opportunity to choose test questions helped her reduce stress. For example, on Notecard #8, she wrote, “We were less stressed because we had options”.

Contrary to what is suggested in calibration research, despite being a high achieving student, it appeared that Neha did not benefit by much from my intervention, which included explicit monitoring strategy training. My qualitative data revealed that possibly, Neha’s self-perception beliefs both as an individual (“I need to be more confident” on Test #4 with an actual mark of 95%), and as a mathematics student (“I was underconfident about my answers” on Test #10 with an actual mark of 97%) might have negatively influenced the development of Neha’s calibration accuracy.

As noted earlier, a student’s miscalibration manifests itself either in the form of an overestimation or an underestimation. My cross-sectional blended data analysis suggested that Neha underestimated her performance outcomes. This is consistent with the perspectives from calibration research (e.g., Hattie, 2013), which states that high-achieving students are underconfident in the sense that they underestimate their performance outcomes. Neha’s cross-sectional blended data suggested that influenced by her self-perceptions, this high-achieving student underestimated her task performance outcomes, time and again.

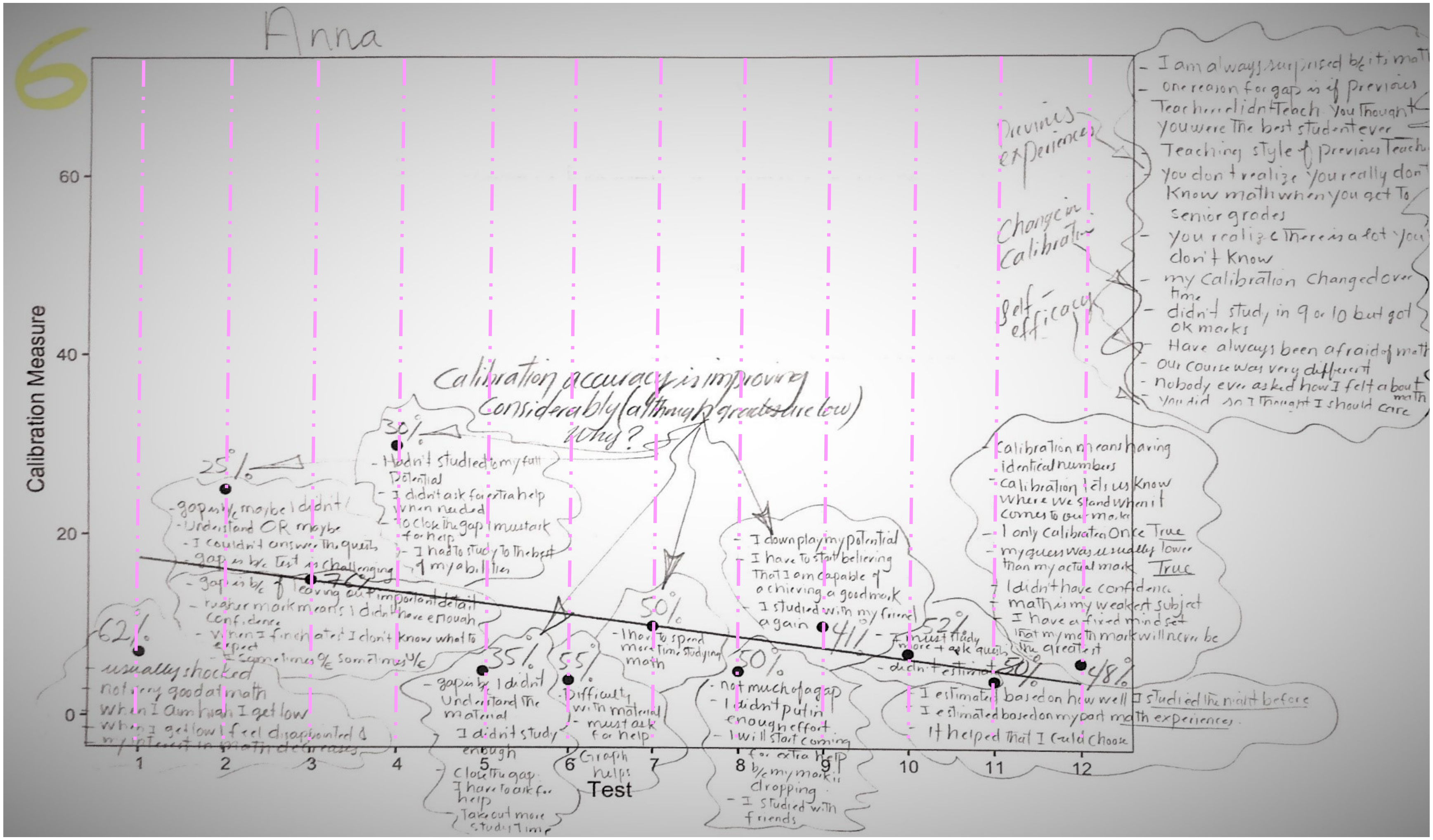
My cross-sectional blended data analysis also revealed Neha’s personal theory of calibration. On Questionnaire #2, in explaining what she thought calibration was, Neha wrote, “Calibration is understanding how your brain works while you work or make decisions. It is understanding why you think the way you do and why and how you make certain decisions." She also added "I think calibration is being aware of your thoughts and questioning them." Although unchartered territory in calibration research, it is possible that how or what students theorise about calibration may affect the development of their calibration, with important implications for calibration research.

The Example of Anna. In a second example of my within-student blended data analysis, I present 12 cross sections of the blended data that I analysed for Anna. In my analysis, I compiled Anna's qualitative and quantitative data, chronologically, on her calibration graph. The issues that influenced Anna's calibration were not the same as the ones that influences Neha's. For instance, on Notecard #5, Anna estimated 30% and achieved 35%. In response to the question of why there was a gap between her estimated and achieved marks Anna wrote, "There is a gap because I did not understand the material, but I didn't study either." Similarly, on Notecard #6 after estimating 50% and achieving 55%, Anna responded that the gap between the estimated and achieved was due to her difficulties with the test material and she should have asked for help. Notably, the small gap of only 5% in both cases indicated Anna's good calibration. Contrary to Kruger and Dunning (1999) although Anna was, possibly, unskilled, she was not unaware of it.

Anna's comments on questionnaires and notecards suggested that in her view, what influenced her calibration were both extrinsic and intrinsic factors. For instance, on Notecard #2 she wrote, "There is a gap because the test was very challenging," giving an extrinsic reason, while on Notecard #12 she commented, "I have a fixed-mindset that my math mark will never be great," providing an intrinsic reason. Figure 4.12 depicts 12 cross sections (illustrated as vertical dashed lines) that I used to analyse Anna's blended qualitative and quantitative data.

Figure 4.12

Anna's Blended Qualitative and Quantitative Data Used in Analysis



On notecards #2 and #12, Anna attributed her miscalibration to her self-confidence beliefs, but on Notecard #9, she commented on her self-efficacy beliefs by stating, “I have to start believing that I am capable of achieving a good mark.” Anna’s other explanations for her miscalibration included not understanding the material (Test #2, Test #5, Test #6), lack of personal effort including not asking for help (Test #4, Test #5, Test #6, Test #7, Test #8, Test #10), and mathematics itself as a school subject-matter (Test #1, Test #12, Questionnaire #1, Questionnaire #2).

Elaborating on the elements of my classroom intervention, on Notecard #6 Anna noted that calibration graphs, which I used for feedback, helped her as she tried to close the gap. On Notecard #11, she noted that it was helpful that she could choose her test questions, referring to my method of using classroom tests as intervention. Additionally, Anna pointed out that her experiences with mathematics, including mathematics teachers were influential. For instance, as shown in Figure 4.12, on Questionnaire #2 Anna wrote, “One reason for the gap is the teaching styles of the previous teachers,” and added, “In [grades] 9 and 10, I didn’t study for math but got OK marks.” On the same questionnaire Anna noted, “Nobody ever asked me how I felt about math, but you did, so I thought I should care.”

Calibration research (e.g., Hattie, 2013) suggests that students who achieve lower grades than their peers are not only less accurate in their estimations, but also less aware that they are inaccurate. Anna provides a counter example in that her calibration improved by 18% over time, while in the case of Neha who achieved much higher grades, calibration improved by only 2%. Furthermore, my blended data analysis suggested that Anna was quite aware of the gap between her estimated and achieved grades. For instance, on Test #8 for which she estimated 55% and achieved 50%, Anna noted, "There is not much of a gap this time, but I did not put in enough effort" and on Questionnaire #2, Anna wrote, "My calibration changed over time."

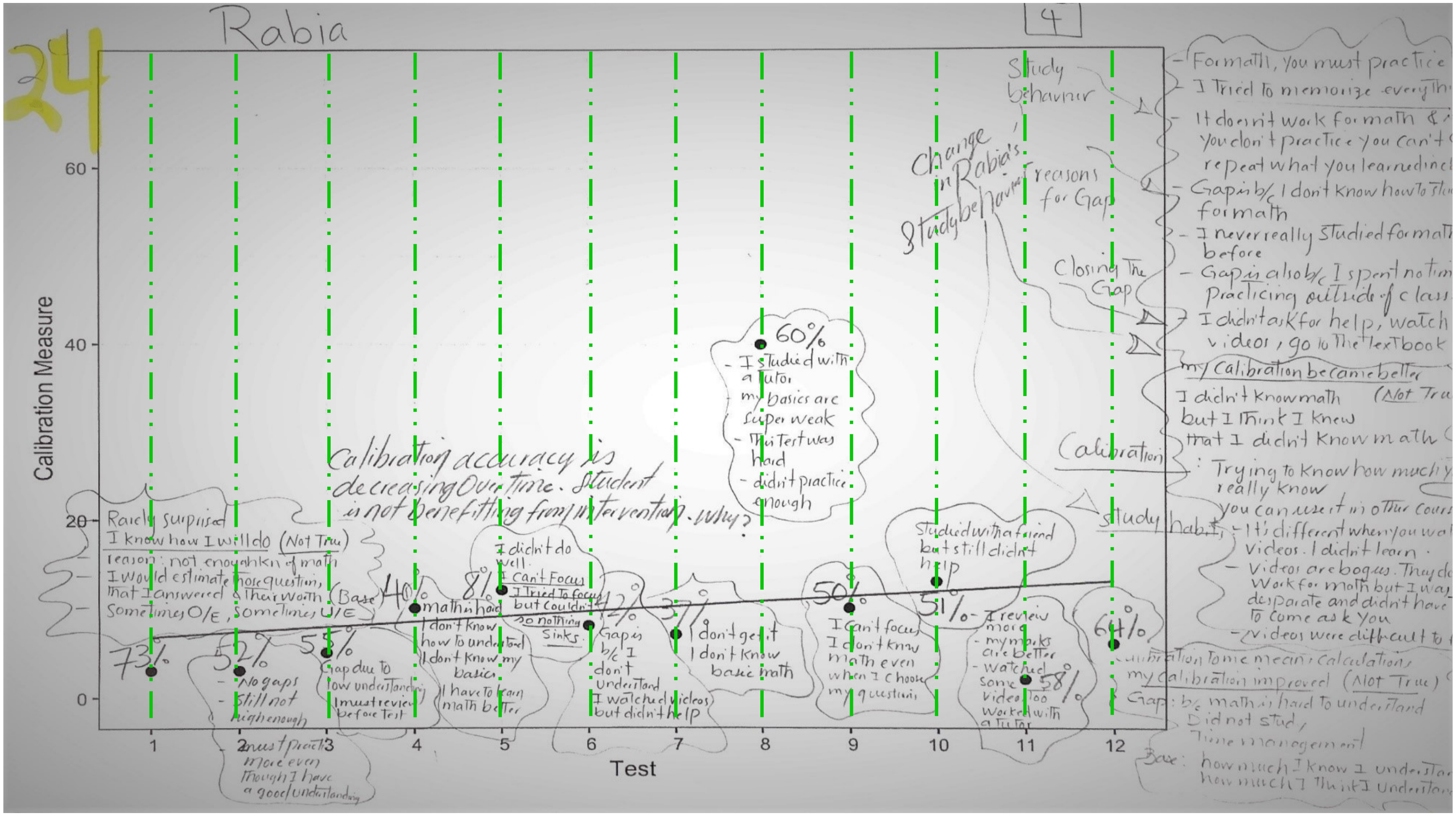
Elaborating on her personal theory of calibration and what calibration meant to her, on Notecard #12, Anna noted, "Calibration lets us know where we stand in terms of our marks." Although I did not specify how and based on what students should make their estimations, Anna did this on her own. For instance, on Notecard #11 she wrote that she made her estimations based on how well she studied for a test. Neha's case represented a high-achieving student for whom the calibration accuracy, as depicted in Figure 4.11, improved by only 2%. However, Anna's case provides an example of a low-achieving student whose calibration accuracy, as illustrated in Figure 4.12, improved by 18%. Both these cases are contrary to the findings of the calibration research that I reviewed.

The Example of Rabia. In the previous two examples, the calibration accuracy improved to varying extents, for both Neha and Anna. However, calibration accuracy did not improve for all the students in the group. My third example is about Rabia whose calibration measures increased over time, indicating that gradually, she became less calibrated. I wondered why my intervention did not benefit Rabia and by analysing Rabia's blended data, I found some possible explanations.

Figure 4.13 depicts 12 cross sections that I used in my within-student blended data analysis for Rabia. The vertical dashed lines illustrate the instances in which I blended and analysed the qualitative and quantitative data for Rabia. The rise of the graph from left to right, indicates a decrease in Rabia's calibration accuracy over time.

Figure 4.13

Rabia's Blended Qualitative and Quantitative Data Used in Analysis



In response to the question of why there was a gap between her estimated and achieved marks, on Notecard #5, Rabia who estimated 20% and achieved 8% wrote, “There is a gap because I didn’t do well. I can’t focus. I tried to focus but couldn’t, so nothing sinks [in].” Rabia seemed to propose that if she could focus, her mark might have been higher than 8%, resulting in a narrower gap between her estimated and achieved marks. On Notecard #8, Rabia estimated 20% and achieved 60%. In explaining why there was a 40% gap Rabia wrote, “I studied with a tutor, but my basics are super weak.” In comparison to Notecard #5, by using her self-regulation (i.e., deciding to study with a tutor), Rabia improved her mark, but the gap of miscalibration did not narrow down, possibly because she did not adjust the level of her estimation to the level of her effort and test preparation. But why? Rabia’s reason took the form of a self-efficacy judgment when she wrote that in her view, her basic mathematics knowledge was weak. It seemed that Rabia’s self-efficacy beliefs affected her estimation and increased the gap. On Questionnaire #2, commenting on the modifications that she made to her study behaviour, Rabia wrote, “I watched math videos, but videos don’t work for math. They were difficult to follow. I didn’t understand them, but I was desperate and didn’t have time to come to you.” These comments suggest that Rabia tried to make changes to her study behaviour, which is an aspect of metacognitive control and related to the process of becoming better calibrated.

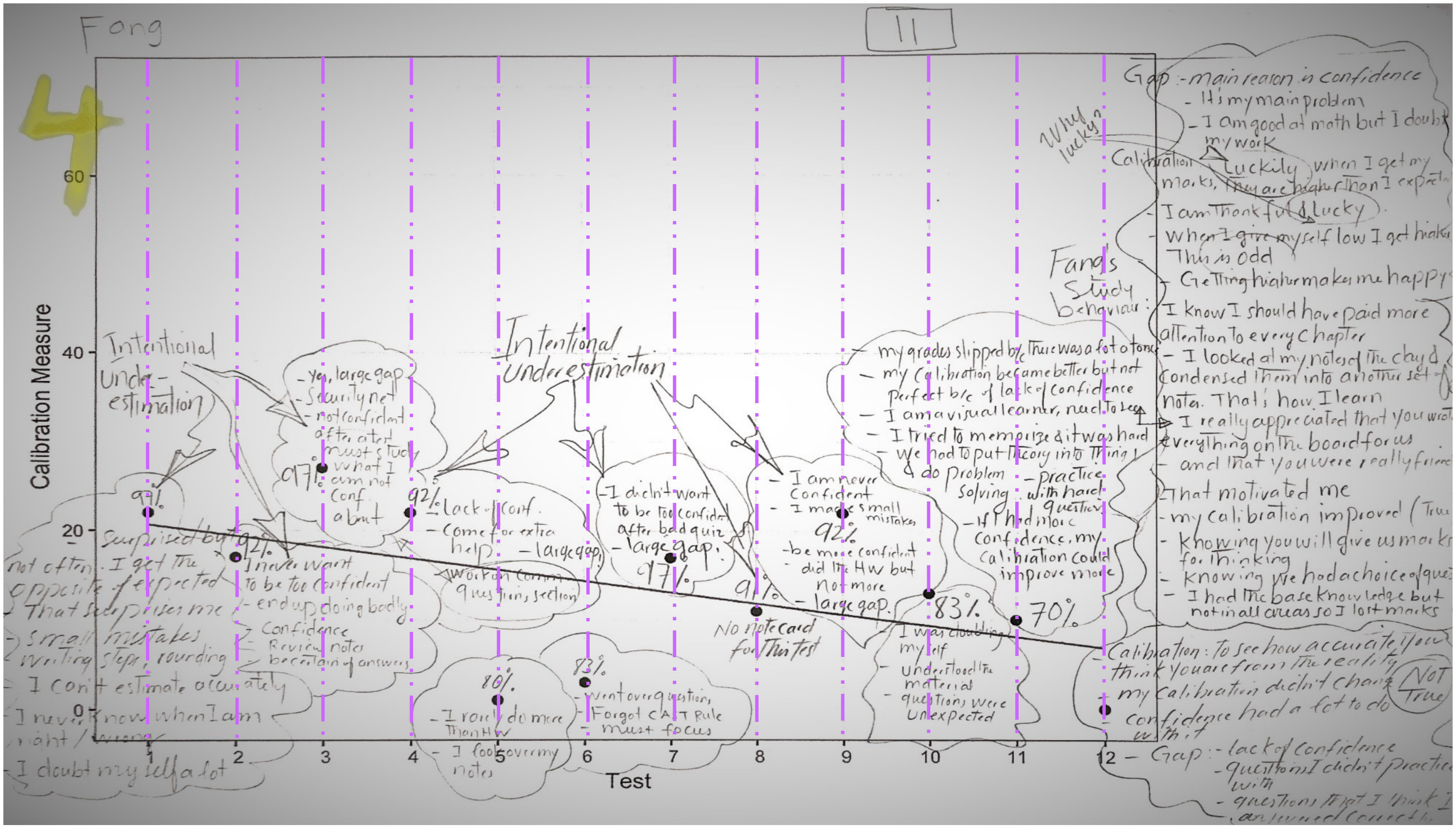
By analysing Rabia’s blended data I was able to explore how she theorised about calibration. On Questionnaire #2, Rabia wrote, “Calibration is really about calculations. You guess your mark and try to match it with your actual mark. It’s all math.” Since I had no access to Rabia’s entire data until after the interviews, I did not realise how limited Rabia’s view of calibration was. Rabia did not connect the notion of calibration with her own personal decisions and self-regulation. Although she made changes to her study behaviour, Rabia understood calibration as a “game of numbers and calculations”, not as an exercise in self-reflection and self-evaluation. Further, on Questionnaire #2, Rabia noted, “My calibration improved over time.” However, in reality, this was not the case, since Rabia’s quantitative data and her individual calibration graph showed that the measures of Rabia’s calibration increased over time, indicating that as time passed by, Rabia became less calibrated. The findings from Rabia’s case are consistent with research findings suggesting that the knowledge of the field (in this case mathematics) might be necessary for making accurate performance estimations in the field (Kruger and Dunning, 1999; Nietfeld and Schraw, 2002).

The Example of Fang. In my final example of within-student blended data analysis, I present the case of Fang. Fang was a high-achieving student whose calibration accuracy improved over time. As noted earlier in the case of Neha, this was consistent with the findings of calibration research (e.g., Hacker et al., 2000), positing that the calibration accuracy of high-achieving students improves over time. However, my within-student analysis in which I compared the same student’s performance outcomes and performance estimations over time, revealed to me that Fang underestimated her performance outcomes 12 out of 12 times. I was intrigued why a high achieving student such as Fang might decide to consistently underestimate her task performance. Calibration researchers often analyse the combined calibration accuracy measures of a group of students. Following the development of the calibration accuracy of an individual student over time and several curriculum-based classroom tests is not common in calibration research. In my review of the research literature, the only example that I could find was Rutherford’s (2017b) study.

Figure 4.14 represents 12 cross sections that I analysed for Fang in my within-student blended data analysis.

Figure 4.14

Fang's Blended Qualitative and Quantitative Data Used in Analysis



In calibration research, to my knowledge, the notion of students' *intentional miscalibration* has not been attended to. This is an important issue because if intentional miscalibration persists, students might not benefit from classroom interventions aimed at developing their calibration accuracy. Additionally, the unrealistically altered values generated by students' intentional miscalibration may potentially prejudice the results of quantitative studies of calibration. It would have been helpful for me as a researcher to have the opportunity to review my data as they were generated to see that despite being a high-achieving student, Fang was intentionally underestimating her performance on every single test.

My within-student blended data analysis suggested that overall, Fang's calibration accuracy improved. I wondered how better calibrated she could have become, had she not intentionally miscalibrated, or had I had access to her data to provide feedback and support.

Moreover, I was intrigued why despite many episodes of feedback Fang continued to underestimate her performance outcomes. Fang's qualitative data provided some possible answers. For example, on Notecard #1, Fang had a 22% gap with an estimation of 75% and achieved mark of 97%. In explaining the reason for this gap, she wrote, "I doubt myself a lot." In a sense, Fang implied that her lack of confidence was the reason for her low estimation leading to a large gap. Similar comments were shared by Fang on her subsequent notecards. On Notecard #2, with a gap of 18% (estimated 74% and achieved 92%) Fang wrote, "I never want to be too confident, because I can end up doing badly." Further, with a gap of 27% on Notecard #3 (estimated 70% and achieved 97%) Fang noted, "There is a large gap, but this gap is my safety net. I am not confident." In analysing Fang's blended data this was the first instant that I suspected her miscalibration was intentional and she had not been trying to close the gap.

In asking students to estimate their performance outcomes *after* writing a test, I assumed that students could accurately judge how they performed due to the testing effect (Schraw, 2009; van Loon et al., 2013). Further, students wrote their notecard comments *after* observing the estimated and achieved outcomes. However, these research design considerations of mine were apparently no remedy for students' intentional miscalibration and my lack of access to my entire data left me in no place to provide constructive feedback for students.

What I Learned From my Within-Student Blended Data Analysis. My within-student blended data aanalysis taught me how individual student's calibration changed and developed throughout the study from their own perspective and also based on the data. It further afforded me a clear perspective on factors influencing the development of each individual student's calibration according to them and highlighted for me the characteristics of an adaptive intervention that might help improve students' calibration.

Considering the aforementioned inconsistencies between the findings of calibration research literature and what actually happened with students in my study, I posit that to create interventions that might benefit students, calibration researchers should consider conducting research in authentic classroom environments with students and listen to what students have to say about their self-perceptions, both as an individual and as a student. I concur with Persky and Dinsmore (2019) who proposed that having students practise their calibration over time and exploring the bases of their self-estimations of task performance has the potential to collect meaningful data on calibration accuracy.

Across-Students Blended Data Analysis

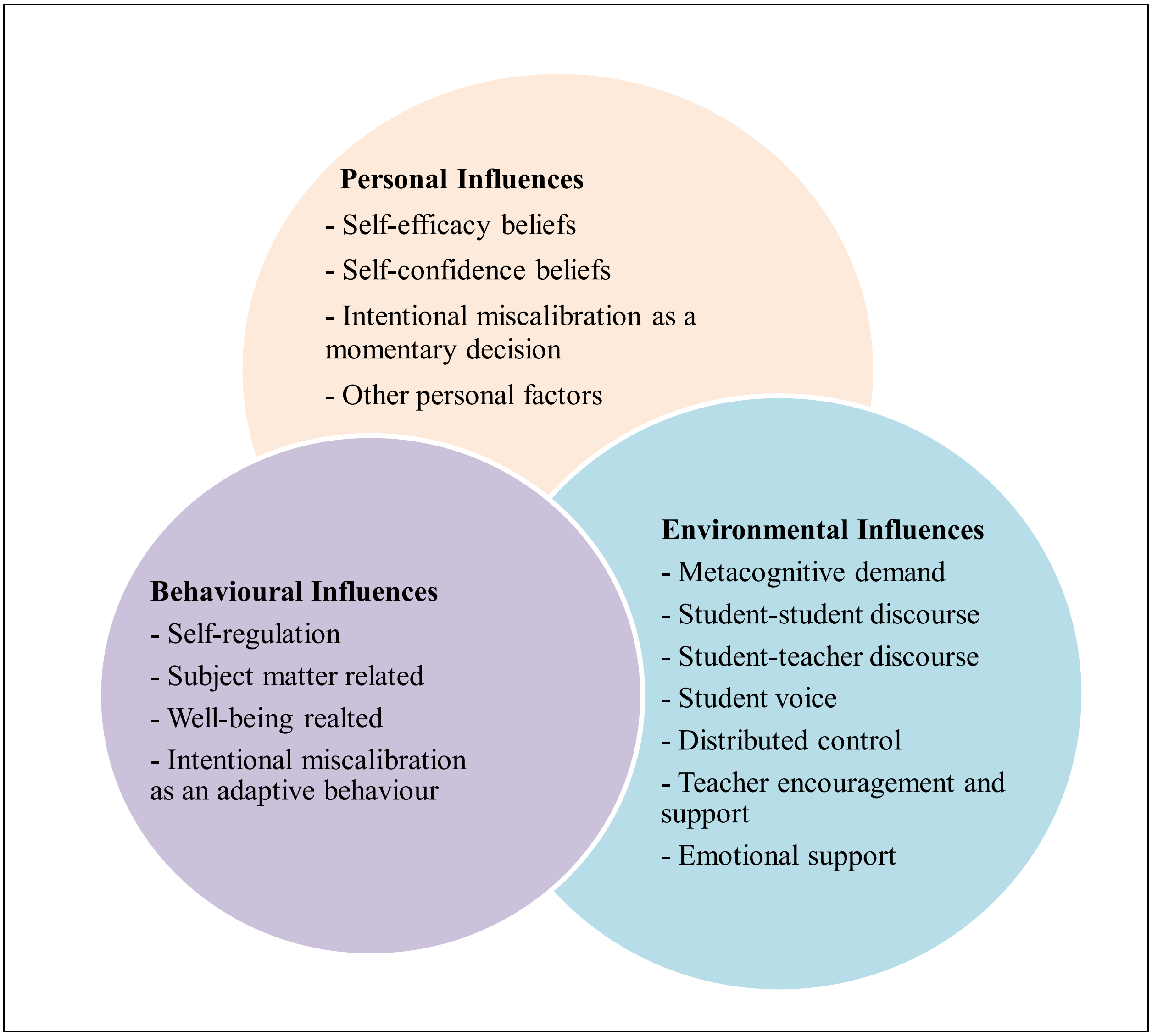
In describing the second round of my blended data analysis, I present what students as a group described in their written comments to be important and influential for their calibration. At this stage, I analysed the entire blended data, except for the interview data, for each individual student and compared and contrasted the results with other students for an in-depth across-students analysis. Consistent with social cognitive theory, I organised the blended data within the triadic influences on students' calibration. In so doing, I explored what students proposed to be the personal, behavioural, and environmental factors that influenced their calibration.

To clarify, from a quantitative perspective, calibration measures are the numerical discrepancies between students' estimations of task performance outcomes and their actual task performance outcomes. From a qualitative perspective, calibration refers to the extent of the soundness of students' monitoring judgments about their own task performance. Working in time-stamped electronic files, I looked for what emerged from students' comments and all the references to personal, behavioural, and environmental perspectives were identified and analysed. As noted earlier, during the intervention phase, when asking students to make estimations after each test, I had not given them any specific criteria to base their estimations on. For instance, I did not ask them to self-estimate based on their previous performance outcomes or based on how well they felt they did on the test. My intention was to keep the data as authentic as possible. The downfall of not suggesting a basis for students' self-estimations was the variability of what students indicated to influence their calibration. I focused on my research question and searched for what students identified consistently and with confidence to have affected their calibration.

Figure 4.15 represents the sub-categories that I generated from the data in my across-students blended data analysis. I chose overlapping circles to depict the reciprocity of the triadic categories of personal, behavioural, and environmental factors, as proposed in Bandura's (1986) social cognitive theory.

Figure 4.15

Influences on Students' Calibration Within Social Cognitive Theory



Students' Views on Personal Influences on Their Calibration. A discussion on personal influences on students' calibration in a mathematics classroom context—for instance, a discussion on the influence of students' self-confidence beliefs, self-reflection, or fear of failure on their calibration—is unchartered territory in calibration research literature. In what follows, I present what students brought to light as the personal influences on their calibration.

Self-Efficacy Beliefs. In social cognitive theory (Bandura, 1997), self-efficacy refers to the "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3). Students identified the influence of their self-efficacy beliefs on their calibration in a variety of ways. For instance, on Notecard #7 Anna stated, "I have to start believing that I am capable to achieve a good mark," and on Notecard #1 Shirin shared, "I doubted myself and didn't think I will do well so I didn't estimate higher." In their written comments, students often mixed the two different notions of self-confidence and self-efficacy. For instance, on Notecard #1 Shirin's comment, "I doubted myself," indicates low self-confidence. While on the same notecard, when Shirin shared, "I didn't think I will do well," she offered a judgment on the outcome of her academic performance, which speaks to her self-efficacy. Further, students identified erroneously positive self-efficacy beliefs to influence their calibration. For instance, on notecard #5 Kristian, who estimated 80% and achieved 62%, noted, "There is a big gap. I stopped studying because I was sure I'd get a good mark on this test." In her comments, Kristian implied that her premature feelings of self-efficacy might have led to the performance outcome that was markedly lower than her estimation of that performance outcome, affecting her calibration accuracy negatively.

Further, my blended data analysis showed that the fear of failure beliefs also influenced students' calibration. For example, on Notecard #2, with an estimation of 10% and actual mark of 60% Adnan wrote, "I estimated low because I was sure I failed," and on Notecard #6, with an estimation of 60% and actual mark of 90% Li Min shared, "I am always afraid maybe my calculations is wrong [*Sic*]."

Self-Confidence Beliefs. According to Bandura (1986), people's confidence in themselves (e.g., "I am good at trigonometry") differs from their confidence in the outcome of their efforts (e.g., "I am going to ace my trigonometry test"). While the former beliefs are aspects of an individual's general self-confidence (i.e., who I am), the latter reflects an individual's self-efficacy (i.e., what I can do). In this study, students' comments attested to the influence of their self-confidence beliefs on their calibration. For instance, Anjani shared, "I always second-guess myself," which indicates a low level of self-confidence. Further, on Notecard #1, with an actual mark of 100% and estimation of 70%, Anjani noted, "I must believe in myself," while on Notecard #10 with an actual mark of 98% and estimation of 80%, Leah wrote, "I estimated low because of my lack of self-confidence."

Intentional Miscalibration as a Momentary Decision. In analysing the blended data, I found indications of students' intentional miscalibration. I was astonished and wondered why students would deliberately miscalibrate their performance throughout the semester. The qualitative data provided insights as to why students might intentionally miscalibrate their estimations as a momentary choice. For instance, after Test #3, Faraz who estimated 60% and achieved 100% noted, "My mood was off on the test. So, I estimated low," implying that his low estimation was related to his mood while writing the test. After Test #7, for which Sheng estimated 40% and achieved 65%, he stated, "I did not feel good about this test, so I estimate low [*Sic*]." In his comment, Sheng shared that he did not feel good about Test #7 and his estimation was, possibly, affected by how he felt about this particular test. In another example, after Test #8, for which Sun Xin estimated 42% and achieved 70%, he wrote, "Not a good test. I decided to estimate low. I was tired." Possibly, Sun Xin's low estimation demonstrates a momentary decision due to being tired at the time, rather than a habitual behaviour.

Other not Previously Identified Factors. Students' comments highlighted other factors not previously reported in research literature. These factors did not repeatedly appear in the data, but they were important to me as I pursued answers to my research question and searched for what might have influenced students' calibration. For instance, on Notecard # 7 Faraz wrote, "I am a pessimist. That's why there is a gap." Further, on Notecard #5, he shared, "My mood was off, and my mood reflects my confidence, so I underestimated." Liu Wei mentioned distraction (Notecard #5), explaining, "I get distracted very easily," while Li Min highlighted the issue of English language efficiency (Notecard #2) by stating, "I don't understand some of the words." Moreover, the students identified the notions of stress ("There is a gap because I was nervous on the test," Marel on Notecard 5) and well-being ("Big gap because I felt too tired and sleepy," Wang on Notecard #6), as the influences on the gap between estimated and achieved performance scores and hence on their calibration.

To summarise, students' comments regarding the personal influences on their calibration accuracy raised not only the issues of self-confidence and self-efficacy, but also brought to light indications of other factors such as mood, attention, anxiety, and well-being. Notably, my analysis suggested that nine students in the group intentionally miscalibrated by consistently underestimating their performance scores. As I noted in the methodology section (see Chapter Three), I did not have access to my entire data while the data were being generated by the students. As such, while some of the participating students continued to intentionally miscalibrate, I had no way of knowing about it, and was completely taken by surprise when eventually I had access to my entire data. Students' intentional miscalibration—which, to my knowledge, has not been explored in intervention studies and is absent from the metacognition research literature in high school mathematics classrooms—deserves attention, as it can effectively bias the results of any purely quantitative studies of calibration.

Students' Views on Behavioural Influences on Their Calibration. In this section, I review students' comments regarding their various learning behaviours (e.g., their study habits, test-taking and test preparation skills, self-regulatory behaviours), and the potential impact of these factors on students' calibration. Questions remain as to how effective classroom interventions might be designed and implemented to promote adaptive change in students' study behaviour so that they can improve their calibration. In this study, one advantage of mixing the qualitative and quantitative data in a chronological way was that it enabled me to look for any indication of adaptive or maladaptive change in students' study behaviour, over time. In the next section, I highlight what students perceived to be important and influential in closing the gap between their performance outcome estimations and their actual performance outcomes to improve calibration.

Self-Regulation. Within metacognition research literature, self-regulation (Zimmerman, 2000, 2002) refers to students' self-generated thoughts, feelings, and actions that are centred on attaining a certain personal goal. The second question on reflection notecards which asked the students how they were planning to close the gap between their estimated and achieved outcomes, intended to inquire about students' self-regulation. As such, students' comments in response to this question reflected the influence of students' self-regulatory decisions on closing the gap of miscalibration. For instance, reflecting on how she could bring her estimated and achieved performance outcomes closer, on Notecard #7 Shirin who estimated 50% and achieved 38% wrote, "I plan to go for extra help and review more to close the gap." Students completed their reflection notecards *after* receiving their marks. Perhaps by realising that her study methods had not produced favourable results, Shirin planned to adjust her self-regulatory methods toward attaining her personal goal of, possibly, closing the gap of miscalibration. In another example, on Notecard #5, Arios who estimated 75% and achieved 46% wrote, “I just scanned my notes and scanning is not studying.” These comments which were made after receiving a lower mark than anticipated, might reflect Arios’s evaluation of his self-regulatory behaviour (i.e., scanning) and his conviction that his self-regulatory behaviour was inadequate.

Interpreting students’ comments is a complex task. However, based on my data, when students like Shirin and Arios who overestimated their performance were asked how they were planning to close the gap, they chose to raise their actual marks to bring their marks closer to their estimations. These and similar comments within the data highlight students’ thinking about changing or adjusting their self-regulatory behaviours to improve their calibration accuracy.

Mathematics as a School Subject-Matter. Students’ comments suggested that mathematics as a school subject-matter could influence their calibration. The mathematics curriculum in Ontario heavily depends on students’ language-based skills (e.g., explanations, conjectures, compare and contrast statements). In their comments, students indicated that their calibration might be influenced in connection to the ways that mathematics was assessed. For instance, Sheng noted, “I need to learn math words [*Sic*],” and Wang shared, “I don’t know how to write English in math [*Sic*].” It is possible that for Sheng and Wang, who were new to Canada, the issues related to English language facility played a role. Beyond language facility, other subject-matter related issues were highlighted by the students. For example, Neha commented on her mathematics-related studying strategies, explaining, “Writing out specific math rules helps me remember them later,” while Andreas wrote about his choice of test questions, stating, “I should not have chosen the CAST Rule questions on the test. I didn’t know the CAST Rule.”

Would the students be better calibrated if even tentatively, the subject of study was anything else other than mathematics? I have no data to answer that question, but my data indicated that students connected mathematics, as a subject of study, to their self-evaluation and calibration.

Well-being Related Issues. Students comments also illustrated well-being related issues which were behaviourally oriented. For instance, Neha wrote, "I pull all-nighters for math," and Liu Wei shared, "I skipped a few math classes and the gap increased after that." In this comment, Liu Wei appeared to hint at not adjusting his self-assessment judgments after he missed a few classes, which is an aspect of calibration. Sheng made a reference to his sleep pattern, "I must sleep earlier and get enough sleep," while Rabia, who had a job, shared, "I need to prioritize my free time if I am going to close the gap." Rabia and Sheng seemed to assume that by prioritizing sleep and time management they could close the gap and improve their calibration accuracy. Had I been able to review the students' comments at the time, I could have advised the students that by strategically adjusting the level of their self-assessments to what they realistically believed to be the level of their performance outcomes, they could improve their calibration.

Intentional Miscalibration as an Adaptive Behavior. In discussing the personal influences on students' calibration, students' intentional miscalibration was discussed as a momentary decision made by some students due to a variety of reasons, such as not being in a good mood on the test, being tired on the test, or not feeling good about the test. Although I found these comments informative, I had no evidence to conclude that these comments reflected students' long-lasting and habitual behaviour. In their comments, however, some students' miscalibration also appeared as an aspect of their adaptive and self-protective behaviour. For instance, on Notecard #7 Bilal who estimated 50% and achieved 75% wrote, "Low expectation keeps you happy in life," and on Notecard #7 Lusine who estimated 60% and achieved 89%

shared, "I always underestimate. The reason is that I just want to be safe in case I got a lower mark." Based on their comments, Bilal and Lusine miscalibrated not as a momentary personal decision but rather as an adaptive self-protective behaviour.

Thomas (2003) posited that a key objective in any effort to improve students' metacognition is to develop them as autonomous, self-regulated learners. The social cognitive theory, which underpins this study, emphasises an agentic perspective (Bandura, 2001), which focuses on empowering students to think about and make adaptive decisions regarding their own academic goals. The blended data analysis suggested that over time, as the students were challenged to become better calibrated, they reflected upon and revised their various study habits. Further, in this section, the notion of intentional miscalibration was brought up again, this time as a type of survival behaviour some students adhered to.

Students' Views on Environmental Influences on Their Calibration. Viewing students as self-reflective and agentic (Bandura, 2001; Thomas, 2003, 2013) rests at the heart of any classroom intervention that aims for adaptive behavioural change. Metacognition researchers (e.g., Bandura, 2001; Thomas, 2003, 2013) suggest that improvements in students' approaches to learning can be stimulated in metacognitively oriented learning environments. Further, based on my teaching experiences, I believe that there is a need for intellectually and emotionally supportive and metacognitively demanding classroom environments. In such an environment, students are encouraged to be aware of and reflective upon how they monitor and guide their own thinking processes and how they might decide to change and improve their thinking processes (e.g., "I find this problem complicated at many levels. I should first try to solve a simplified version of this problem with fewer variables").

As I noted earlier in Chapter Three, I used the model of Metacognitively Oriented Learning Environments (Thomas, 2003) to inform my classroom environment and the intervention that I employed to support the development of students' metacognitive calibration. The seven aforementioned dimensions of the model were initially developed for high school science classrooms. This is possibly the first study of metacognition which uses these dimensions in a high school mathematics classroom.

With respect to establishing a metacognitively oriented classroom environment, the students' assertions reflected the changes that they identified in their classroom environment. I discussed the influences of a metacognitively demanding classroom environment in the development of students' calibration earlier on in the literature review (see Chapter Two). Although as noted by Thomas (2003), interpreting students' assertions is neither easy nor straight forward, in the following examples I describe the classroom environment of this study based on students' assertions. To demonstrate which dimensions of Thomas's (2003) model of Metacognitively Oriented Learning Environments, or the model, were supported by students' comments, I have placed students' comments within the seven dimensions of this model.

Example #1: Metacognitive Demand. An aspect of the intervention utilised in my study was the use of calibration graphs as feedback. As such, students received both performance feedback (i.e., how am I doing) and calibration feedback (i.e., how far away am I from perfect calibration). In their comments, students referred to the utility of calibration graphs in informing their calibration and mathematics learning. For instance, on Questionnaire #2 Neha noted, "I feel that my calibration has improved as I am now able to understand what I am doing and why." She also added, "You kept asking us to do those calibration graphs in class and I can now recognize what my weak points are and what I struggle with. I slowly started to close the gap between what

I knew and what I thought I knew." Neha was correct; as depicted in her overall calibration graph (see Appendix K), her calibration improved over time. Neha's comments reflect metacognitive demands, the first dimension of the model that she felt in the classroom environment.

Example #2: Student-Student Discourse. In this study, students could choose to work individually or in small groups. In this regard, on Notecard #4, Andreas wrote, "I always do everything better when I work with my friends in class," and on Notecards #6 and #10, Neha shared, "In class I work with Anjani and it helps me realise the gap between what I think I know and what I really know." These and similar comments speak to the social dimension of metacognition and reflect the student-student discourse, the second dimension of the model (Thomas, 2003).

Example #3: Metacognitive Demand, Student-Student Discourse, and Student-Teacher Discourse. In a metacognitively demanding environment, students must be aware of their learning processes. Within the intervention used in this study, while on problem-solving tasks, I asked students to explain—to me and to their groups—*what* they were doing, *why* they were doing it, and *how* it would help solve the problem at hand. This style of strategy training placed a metacognitive demand on students and supported their calibration. In this respect, on Questionnaire #2 Leah wrote, "I learned better because you asked us to explain what we were doing to you and to each other. It made me think about what I was doing and why I was doing it. I could not lose track. After a while, I was doing it even in physics." Leah's comments speak to the metacognitive demand (the first dimension), student-student discourse (the second dimension), and student-teacher discourse (the third dimension) of the model (Thomas, 2003).

Example #4: Metacognitive Demand, Student-Teacher Discourse, and Teacher Encouragement and Support. In their comments, students clearly differentiated between the classroom environment in this study and their previous classroom environments. For example, on Questionnaire #2 Abrar wrote, “The different thing was calibration. We didn’t do any of that stuff in the previous math courses. The difference was that when you taught us how to use our calibration and helped us graph it, I started to realize my strengths and weaknesses.” Abrar’s comments relate the metacognitive demand, student-teacher discourse, and teacher encouragement and support dimensions of Thomas’s (2003) model.

Example #5: Distributed Control. A key objective in developing students’ metacognition is to develop learners as agentic, self-regulated (Bandura, 2001; Thomas, 2003) individuals. The fifth dimension of the model (Thomas, 2003) recommends the distributed control of the learning and assessment tasks. In this study as noted earlier, on classroom tests students did not have to answer all the test questions. Instead, I provided a pool of pertinent test questions for students to choose from, and students selected and answered the questions that they were confident answering. For example, on Notecard #12 Yasir said, “Students were comfortable knowing that there are options, like they could choose their own questions on a test. They knew there was a way out.” By exercising distributed control over the classroom tests, my main intention was to encourage the students to think carefully about which questions to choose, and to differentiate between what they had learned and had not learned yet. For students like Yasir, this method also provided some level of comfort.

Example #6. Student-Teacher Discourse, Teacher Encouragement and Support, Emotional Support, and Student Voice. Thomas (2003) attests to the importance of considering affective and support issues in relation to developing metacognitively oriented classroom environments. In this study, the affective influences on calibration emerged from students' comments. For example, on Questionnaire #2, Leah wrote, "You need a math teacher who is not going to get mad if you ask the same question over and over again. I really appreciated, we all did, that you were so patient with us and so friendly." Moreover, on Questionnaire #2, Anna commented, "Math has never been my forte. It makes me nervous. I still think that I could have done better this time because I could get help from my teacher and my friends." Further, Anna reflected, "Our course was actually different because the first thing that you asked us was, 'OK, how do you feel about math? How can I make it better for you?' I was like, if my teacher cares so much, maybe I should care a little more." Leah's and Anna's comments relate the student-teacher discourse (the third dimension), student voice (the fourth dimension), teacher encouragement and support (the sixth dimension), and the emotional support (the seventh dimension) of the Thomas's (2003) model.

What I Learned From my Across-Students Blended Data Analysis. Based on the dimensions of Thomas's (2003) model, the classroom environment in my study seemed to have stimulated the development of students' calibration; a notion supported by students' assertions. To summarise, by blending and analysing the quantitative and qualitative data, I was able to form constructive assertions regarding what influenced and changed students' calibration personally, behaviourally, and environmentally. For instance, the quantitative data showed that the overall measures of calibration decreased over time, suggesting that as a group, students became better calibrated. The qualitative data highlighted the notion of intentional miscalibration that nine students consistently exercised. My pedagogical takeaway was that notwithstanding what I do, say, or teach, students make their own decisions, and a teacher's intervention efforts do not always inspire adaptive change on the part of the students. I also learned that as evidenced by the quantitative data, despite the intentional miscalibration of more than one third of the students, calibration accuracy improved collectively over time, attesting to the robustness and multidimensionality of the construct of calibration.

Incorporating Students' Interview Data

Before adding the interview data, I was concerned about the consistency of my data in asking the students the same question using different research methods. For instance, I wondered if the students had given me the same reasons for their underestimations in their written notecards and questionnaires as they had offered during their interviews. Hence, in this third and final round of my data analysis, I revisited the factors which students had previously identified in the blended data as being influential for their calibration and compared and contrasted those factors with the qualitative data from students' follow-up interviews.

I conducted the follow-up interviews in the post-intervention phase after the end of the semester. Therefore, there was a time gap between students' written and verbal assertions. Listening to the students' explanations about their views on calibration and their perceived reasons for miscalibration, I focused on what they shared with me without the constraints of written communication, as the students could speak openly and extensively about their personal experiences with calibration, my intervention, and the learning environment. The use of follow-up interviews to determine whether students would confirm or disconfirm in spoken language what they have asserted in written form is a research practice which to my knowledge has not been reported in calibration research literature.

During the interview sessions, students spoke to me about their experiences in our classroom environment, highlighting issues that they deemed important. As a researcher, I was pleased to have the opportunity to compare and contrast what students related in their interviews with what they had already written on their questionnaires and reflection notecards.

All students were invited to participate in the follow-up interviews, which were conducted after the end of the school year during the first two weeks of July 2016. However, at that time, some students had returned to their home countries. Twelve students accepted my invitation and participated in the follow-up interviews. Consistent with the previous intervention related activities, in the context of the follow-up interviews, the students used the term gap, possibly, to refer to the discrepancy between their estimated and achieved grades on classroom tests and used the phrase ‘closing the gap’ synonymous with ‘becoming better calibrated’.

The Example of Shirin

I present the example of Shirin because she was the only student in the group who said in her interview that she felt *embarrassed* by her calibration graphs. This was surprising for me, because I did not expect such an emotional response to the presumably simple act of point plotting on a graph. Shirin also shared that her calibration graph made her more realistic about her performance outcomes in mathematics.

Over the course of the intervention, Shirin related her miscalibration to her overconfidence. On Questionnaire #1, she wrote, “Most of the times I am surprised by my math marks. I tend to get a lower mark whenever I feel I have actually studied,” and on Questionnaire #2, end of the semester, she noted, “I think the main reason for the gap was my overconfidence in my knowledge and then studying less, because I thought I knew what I was doing.” In her interview, Shirin confirmed her comments by stating, “Most often, I study for math the night before. I start studying and thinking I must get a really high mark. My calibration graph was really embarrassing for me, because sometimes I would estimate a lot higher than it actually was, so I was like…I need to close the gap.” Further, in her interview Shirin shared, “My calibration graph helped me become more realistic about my schoolwork.” Shirin’s interview data confirmed what she had written earlier on questionnaires and notecards. Although I had no access to my entire data until after the interviews were completed and I worried about the possible inconsistencies in what students wrote throughout the intervention with the interview data, there was consistency in the data.

The Example of Andreas

In a second example of the congruency of students’ comments during and post the intervention phase, on Questionnaire #1, Andreas shared, “I expect a low mark and keep my expectations low.” On Notecard #5, he noted, “There is a gap because I do not expect a high mark. I have always been afraid of math.” On Questionnaire #2, Andreas stated, “Math is not like other subjects for me. I am afraid of math. The gap is there because I am afraid and have no confidence when it comes to math.” In his interview, Andreas confirmed, “I have been surprised by my math marks so many times. Also, I don’t have a good base for math…so I kept estimating very low because I am really scared of math. I have always been afraid of math.” Andreas comments during the follow-up interview confirmed what he had written before, throughout the semester, and it became clear to me that Andreas had underestimated his performance scores based on his personal feelings about mathematics. This issue has implications for teachers due to the potential maladaptive influences of students’ feelings about mathematics on their motivation and achievement.

There are implications for calibration researchers who might not be able to design the research studies that work for students due to ethics considerations. Had I been able to review my qualitative data as they were being generated during the semester, I would have suggested a basis for students' performance estimations. For instance, I could have advised Andreas that despite his self-proclaimed fear of mathematics, his performance estimations should be based on the quality of his review and test preparation studies.

The Example of Zishan

In another example of the possible congruency of students' written and interview comments, on Notecard #4, after observing a mark of 100%, Zishan who had estimated 100% wrote, "No gap! I rewrote all my class notes and redid all the homework to close the gap," and on Notecard #6, after estimating 100% and achieving 92%, Zishan shared, "I have to go back to the strategy that I had when I got 100%," indicating that he intended to adjust his study behaviour to close the gap. In his interview, confirming the adaptive changes that he made to his study behaviour, Zishan said, "We don't always know what we know and what we are going to get, but we have that feeling of what we know, what we have to get, and how we want to get it. I changed how I studied for math to close the gap."

In his interview, Zishan further shared, "[To close the gap] I redid all my class notes. I actually rewrote all my class notes, because in class, I was rushed, and the notes were not very clear…I did everything one more time, and it worked." To provide context, over the course of the semester, Zishan had estimated very highly, predominantly 100% to motivate himself. He had then tried to close the gap between his estimated and achieved marks by rewriting his class notes, which he hoped would help improve his marks. As confirmed by his interview comments, this method worked well for Zishan and enabled him to narrow down the gap of miscalibration.

The Example of Neha

To provide a fourth example of the consistency in the data, I present the case of Neha whom I introduced earlier in this document. On Notecard #1, Neha stated, "I felt underconfident in myself. I did not believe I could do extremely well. I did not have self-confidence, hence causing the gap" (her estimation was 80% and her actual mark was 97%), and on Notecard #4, she stated, "The gap is due to my lack of confidence, because I was very stressed during the test and thought I would fail" (her estimation was 70% and her actual mark was 95%). Further, on Notecard #6, Neha wrote, "The gap is due to lack of confidence and careless mistakes. I thought I should have studied harder" (her estimation was 60% and her actual mark was 98%). Toward the end of the semester, on Notecard #10, Neha shared, "I was underconfident about my answers and also thought I forgot how to do stuff. I did not know if I was doing it right and I did a lot more work than I needed" (her estimation was 70% and her actual mark was 97%). Overall, throughout the intervention, Neha pointed out that her calibration was influenced by her self-confidence as well as her self-efficacy beliefs.

In her interview, Neha continued to bring up the issue of confidence as the most influential factor in her calibration, stating, "I am very underconfident with my answers and second-guess myself a lot." Although I welcomed this consistency, I realised that despite improvement in Neha's overall calibration (see Appendix K), Neha's self-perceptions remained predominantly intact. I was thus reminded that for me as a teacher and researcher, focusing on what students are confident about might be as important as focusing on the development of their calibration.

As noted earlier, 12 students participated in the follow-up interviews. In addition to the aforementioned four examples, my analysis showed that the other eight students' written comments were also confirmed by their verbal assertions during the interviews. Thus, in the third round of my analysis, the intervention's notecard and questionnaire assertions for students who participated in interviews were consistently matched by their interview statements. The consistency of what students revealed during their interviews and their previous comments was reassuring regarding the authenticity of the data in that the research data reflected students' genuine thoughts and feelings and that there was consistency over time. This issue has implications for methodology and pedagogy in relation to classroom interventions that are designed to improve calibration accuracy.

Overall, by conducting three rounds of blended data analysis, I followed the change in students' calibration accuracy, as well as the evolution of students' perspectives on calibration. In so doing, I sought to explore potential answers to my research question: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom? In the next section, I revisit students' perspectives on what they perceived to have influenced their calibration and how they thought their calibration changed in the context of the intervention. Due to the enormity of the data, in what follows I present the students' perspectives under seven assertions made based on what was categorically consistent in the three rounds of the blended data analysis.

Assertion #1: Overall, students placed a high value on personal study strategies such as notetaking, reviewing, and completing homework, in relation to the change in their calibration. As noted earlier, on the reflection notecards which were repeated every week, I asked the group why there was a gap between their estimated and achieved marks (if such a gap existed) and how they were planning to close this gap. To clarify, in the absence of any other context or frame of reference, in my analysis of students' comments in response to these two questions, my interpretation was that by using the term 'gap' students referred to the discrepancy between their estimated and achieved marks, in the same way that I had taught them the concept and in the same manner that we had used the term in our classroom conversations.

In their responses as to how they were planning to close the gap, students placed a high value on their study efforts. For instance, on Notecard #6, Adnan who estimated 85% and achieved 67% shared, “To close the gap, I have to start properly studying and doing all the homework.” It seemed that Adnan perceived his calibration as a function of his achievement and assumed that better marks meant better calibration. It made sense for students like Adnan who overestimated their task performance to plan to raise their marks by making adaptive adjustments to their study behaviour and hence narrowing the gap between their estimated and achieved marks. However, in his interview, Adnan provided a different perspective by stating, “What I really tried to do [to close the gap] was to fix my work habits. By doing that I hoped that I would feel more confident, and like…I would close the gap.” This was informative for me because in these comments, Adnan related his calibration accuracy to his confidence.

Some students shared the view that to become better calibrated, they had to increase their study efforts, but they did not comment on the notion of confidence. For instance, on Notecard #2 regarding how to close the gap, Leah wrote, “To close the gap, I feel like I need to study more,” Li Min noted, “to close the gap [I should] practice more to know the words [*Sic*],” and Rabia shared, “To close the gap, I must first practice more to understand the material.”

As a teacher, I am undoubtedly supportive of students’ study efforts. However, what I hoped for students was to adjust their performance estimations to the level of their understanding and preparedness (Lichtenstein et al., 1982) to improve their calibration. As I discussed in my blended data analysis, it is possible for low-achieving students to possess good calibration accuracy and their calibration accuracy may improve over time, as was the case for Anna. Alas, as noted earlier, I had no access to students’ notecard comments until after the interviews and was unable to provide feedback and guide them in the right direction.

Some comments reflected students' self-perceptions in connection with their study behaviour. For instance, on Notecard #2, when asked how he was planning to close the gap of miscalibration, Sheng gave credence to personal study efforts, "I need to study harder, read the textbook, do more work." However, on Questionnaire #2, which he completed toward the end of the semester, Sheng included a self-perception belief when he related, "I think my calibration is bad because I am lazy. I didn't work hard."

Some students placed a high value on directing their study efforts toward improving their English language facility to close the gap of miscalibration. For instance, on Notecard #3, Zhang wrote, "[There is a gap because] I not understand the meaning of some words [*Sic*]." Later, on Questionnaire #2, he shared, "I not understand the words. I write slow. I have to study words [*Sic*]". Similarly, Marel noted, "There is a gap because I just arrived from … and did it all in Arabic so I have to improve my mathematics English [*Sic*]." To my knowledge, the notion that students' language facility might affect their calibration has not appeared within the research literature. As an aspect of the due care for students, this issue deserves attention with respect to any language of instruction used in calibration research.

Students' decisions regarding doing more homework, reviewing their class notes, practicing with more complex questions, asking for extra help, and watching educational videos are components of their self-regulation. In the context of my classroom intervention, as students became aware of the gap between their estimated and achieved performance outcomes, some students focused predominantly on improving their task performance outcomes to close the gap of miscalibration, while others decided to work strategically to improve their calibration. Overall, the blended data encompassing the entire quantitative and qualitative data, consistently suggested that students related their calibration accuracy to their study habits and efforts.

Assertion #2: Some students intentionally miscalibrated despite several episodes of reflection and feedback. As discussed earlier, my blended data showed that some students intentionally miscalibrated. For instance, Anjani's comments attested to her intentional miscalibration throughout the semester. On Questionnaire #1, Anjani wrote, "I usually don't estimate accurately and estimate lower because I don't want to set high expectations and end up disappointed." On Notecard #1, after estimating 75% and achieving 100%, Anjani shared, "I always underestimate just in case I did badly." Later, on Notecard #10 with an estimation of 70% and an actual mark of 97%, Anjani illustrated her continued intentional miscalibration by stating, "I was underconfident about my answers, so I decided to underestimate to be safe." Finally, on Questionnaire #2 Anjani noted, "I underestimate my knowledge because I am scared of disappointment." In her interview, Anjani's confirmed her previous comments:

> Some students like me may think that they may do badly, but they end up getting a high mark because they studied. I set the expectation really low for myself. That's why there is such a gap. I think when I saw the gap, I was happy because I was having low expectations, but I was getting a high mark. That is how I will personally do it. When I set a low goal and get a high mark, I am kind of happy and in a way, I am happy about the gap. For me the gap was really intentional.

Anjani's written and spoken comments confirmed that she had intentionally kept her self-estimations low for self-protection and motivation. In so doing, for a variety of reasons, Anjani intentionally underestimated her performance outcomes, which affected her calibration.

To provide another example, on Notecard #1, Fang noted, "I never want to be too confident. I always underestimate. It is my safety net". Comments regarding students' intentional miscalibration, which surprised me the first time that I was working through the blended data set, were confirmed by students' open and elaborate assertions during the interviews. For instance, on Notecard #2, Fang shared, "I never want to guess something too high and get my hopes up. My estimation is kind of like my safety net," and on Notecard #10 Fang commented, "I am never confident. I thought I did the Knowledge section all wrong. I underestimated to be safe". Fang's comment regarding the Knowledge section on the test speaks to the development of Fang's calibration and demonstrates the increasing sophistication in her metacognitive knowledge. During the interview with Fang, I had the opportunity to learn more about the reasons behind her intentional miscalibration:

> I think the most important reason for the gap is confidence. For me, confidence was the main problem. I know I am good at math. I do not want to gloat, but I do think I have math skills. However, when I am doing some of the questions, I am like, OK is this right? Then I am not sure anymore. I start thinking, OK, I will give myself 75% to be on the safe side. Luckily, when I get my grades back, they are always so much better than I expected. I am lucky. Getting higher than what I expected always makes me happy…

On the positive side, Fang's interview comments were consistent with what she had already shared in written form. It made me realise, however, how resistant to change her self-perceptions were, since these self-perceptions had barely changed after several episodes of strategy training, reflection, and feedback.

Assertion #3: Supporting students' confidence for the development of their calibration deserves attention. As I noted earlier, in the context of this study, students used the two phrases of 'to close the gap' and 'to become better calibrated' synonymously. With respect to the notion of closing the gap of miscalibration, the references to students' self-perceptions appeared across the data. For instance, on Notecard #6 Adnan wrote, "There is a gap because I was not confident about some questions." On Notecard #4 Tahir shared, "There is a gap because I have no confidence," and on Notecard #5 Andreas noted, "I have to become more confident to close the gap." Throughout the semester, I was unaware of what students wrote on their questionnaires and notecards. When I started to read through the data, I realised that students had used some of their own terminology and ways of explaining phenomenon to answer my questions on the questionnaires and notecards. For example, although in teaching about calibration I did not use the terms 'overconfidence' or 'underconfidence,' and strictly talked and taught about 'overestimation' and 'underestimation,' as reflected in students' assertions, they adhered to the notions of overconfidence and underconfidence, and used this confidence-oriented terminology side-by-side with my estimation-oriented terminology.

Assertion #4: Students' mathematics-related experiences, including performance outcomes, classroom teachers, and mathematics-related stress were identified by students as factors that influenced their calibration. In this regard, on Notecard #6 Anna wrote, "I estimated based on my past experiences in other math courses," and on Notecard #4 Bilal noted, "I overestimated myself because if I put my effort in, I get a higher mark in other classes." In another example, on Notecard #12 Abrar noted, "I think I was never right about my mark, because it is math. I didn't believe I could get a high mark." On Notecard #10 Kristian shared, "I want to have high marks so I want to put high expectations, but in reality, I can't reach those expectations, because it is math." These comments suggest that in the future studies on calibration, the role of mathematics itself as a subject-matter deserves attention.

With regards to the role of mathematics teachers in influencing students' calibration, on Questionnaire #2 Fang wrote, "We all appreciated that you were really friendly… I think that's what motivated me, and my calibration improved over time." Reflecting on mathematics related stress and anxiety, on Notecard #8 Anna wrote, “I underestimated because math is my weakest subject. I am always nervous. I also have a fixed mindset that my math will never be the greatest.” In her interview, Anna confirmed the anxiety that she felt for the subject-matter, stating, “I think it is difficult for me because I have always been afraid of math…math had never been my forte,” and also added another dimension by stating, “The reason for the gap maybe depends on who your past teachers were.” The topic of mathematics-related anxiety and its relation to metacognition has been attended to by metacognition researchers (e.g., Legg & Locker, 2009). To my knowledge, however, the influence of such anxieties on students’ calibration has not yet been explored.

Assertion #5: Students identified the four components of my classroom intervention to have influenced their calibration. Students’ references to the four dimensions of my classroom intervention—strategy training, classroom tests, feedback, and reflection—emerged from the data as the influencing factors on students’ calibration.

The first component of my classroom intervention was monitoring strategy training. In this regard, on Questionnaire #2 Leah noted, “I learned better because you asked us to explain what we were doing, to you and to each other. It made me think about what I was doing and why I was doing it. I could not lose track. After a while, I was doing it even in physics.”

The second component of my intervention relied on curriculum-based classroom tests in which students were given a metacognitively oriented opportunity to choose their test questions. In this regard, on Questionnaire #2 Yasir noted, “Students were comfortable knowing that there are options, like they could choose their own questions on a test. They knew there was a way out,” while Shirin shared, “This time there is less of a gap because you always let us choose our own questions and I think I am getting better at it.”

The third component of my intervention was providing feedback by using calibration graphs. In this regard, on Questionnaire #2 Neha shared, “I feel that my calibration has improved as I am now able to understand what I am doing and why. You kept asking us to do those calibration graphs in class and I can now recognize what my weak points are and what I struggle with. I slowly started to close the gap between what I knew and what I thought I knew.” On Questionnaire #2 Kristian wrote, “We did those graphs and it reminded us and motivated us,” while in her interview, Shirin shared, “My estimates were what I thought I knew and then, wow! I didn’t know any of that. My calibration graph helped me become more realistic about my schoolwork. It was an eyeopener.”

Finally, the fourth component of my intervention, reflection notecards, was mentioned in students’ assertions as “did not let us forget about our calibration” (Anna’s interview), and “notecards were like a mirror. They showed us where we were” (Yasir’s interview). In these comments, students relayed what they viewed to be useful about my classroom intervention and how the four elements of strategy training, classroom test, feedback, and reflection influenced their calibration.

Assertion #6: Students described the adaptive, self-generated changes that they made to their study habits in the context of this classroom intervention. On Notecard #1, Anjani noted, “To close the gap, I must continue to practice for math everyday.” On Notecard #3, regarding her study behaviour, Anjani wrote, “I should practice more on complex questions rather than doing everything.” Later, on Notecard #9, she made more changes to the ways that she studied mathematics by adding a study partner, “I must redo all the homework questions, work harder, and work with Neha.” In her interview, Anjani shared, “To close the gap, I tried to go to the textbook, ask for help, and watch videos," indicating that she continued to change and modify her study behaviour.

The adaptive, self-driven changes that students make to their study behaviour are aspects of their self-regulation because as noted by Zimmerman (2002) these changes take place when students use self-generated thoughts, feelings, and strategies to attain their study goals. As such, adaptive change to students' study behaviour was a pedagogical goal of my intervention, which intended to support the students in becoming metacognitive, self-regulated learners.

Assertion #7: Over time, students developed their own personal theories of calibration which indicates a change in students' metacognitive knowledge. In her interview, Neha shared, "From all the lessons I have understood that calibration is about understanding how you think and how you make decisions. We usually make judgment calls and make decisions. Calibration means questioning those decisions. Calibration questions why you are doing this and why you are doing that." In these comments, Neha elaborated on her personal theory of calibration indicating a change in her declarative metacognitive knowledge which I discussed in detail in Chapter Two of this thesis. Students' personal conceptualisations (i.e., theories) of calibration—which, to my knowledge, have not been explored in calibration research—highlighted the ways in which students theorised about the construct of calibration and what it entailed. For instance, during the student interviews, Shirin explained, "Calibration really shows what you think you are doing as opposed to what you are actually doing. It is like you recognize there are things that you can do so that you understand math better," while Yasir noted, "Calibration is like a fitness course, you know. You take a photo of yourself to see how you did before and after…you flex your math muscles!" Shirin and Yasir's comments also attest to the change in these students' metacognitive knowledge.

Relating her understanding of what she thought calibration was, in her interview, Fang commented, “We talked about calibration a lot and I know that it is very helpful in terms of what you need to work on and what you need to work less on. It doesn’t need to be in a classroom. You can use it in life in general.” In her later interview comments, Fang shared, “Calibration teaches you that you have to set goals for yourself if you want to be successful. In anything you have difficulty in life, you can set a goal and work towards it, but you have to work.” Fang’s comments indicate change in her declarative, procedural, and conditional metacognitive knowledge.

Other students also offered personal theories about what they thought calibration was. Some of these comments were in connection with students’ knowledge and understanding of mathematics. For instance, Kristian wrote, “Calibration is about knowing what you know and don’t know. I realized I did not know as much math as I thought I knew,” and Anjani said, “Calibration helps you prioritize your time on learning what you don’t know over what you already know.” Students also connected calibration to their self-regulation. For instance, Adnan shared, “Calibration is all about questioning the decisions that you make about how you study,” and Marel wrote, “Before this course, I never thought about setting a goal and trying to get there. Now I do that in all my courses.” Notably, these comments attest to the change and development in students’ metacognitive knowledge. For example, when Faraz shared, “Calibration made me realize how much math I did not know and how I could improve my work by asking those what, why, how questions that you taught us,” he demonstrated the change in his declarative and procedural metacognitive knowledge.

The results of my analysis suggest that in students’ views, the relationship between one’s academic task performance and calibration was bi-directional. To explain, students suggested that not only improved task performance could enhance one's calibration, but also improving one's calibration could potentially lead to better task performance outcomes. Calibration researchers have attended to this issue only unidirectionally (Bol et al, 2005; Hacker et al, 2000; Isaacson & Fujita, 2006; Jacobson, 1990; Maki, 1998; Miller & Geraci, 2011), suggesting that high-achieving students are more accurate in their calibration judgements than their low-achieving peers. The students' comments on the reciprocity of this relationship suggested that they believed that to achieve a high level of performance, they needed to become better calibrated than they already were. For instance, on Questionnaire #2, Leah wrote, "The more calibrated you are, the better you will be able to study," and during her interview, Shirin shared, "I think calibration helps you become more aware of your knowledge level and what your strengths and weaknesses are. It can improve your marks." In these comments, students suggested that better calibration might improve their task performance outcomes. On the other hand, on Questionnaire #2, Li Min noted, "If you can get good marks, then you know your level and you know good calibration [*Sic*]," and Faraz wrote, "Because calibration is the ability to know your strengths and weaknesses, if you want good calibration first you need good math marks." In these comments, students posited that improved task performance might lead to enhanced calibration, hence, suggesting the reciprocity of the relationship between calibration accuracy and task performance in mathematics.

During the follow-up interviews, students elaborated on the utility of calibration beyond academics. In a sense, they generalised and transformed what they had learned about calibration in the classroom context to a much broader social context, which I found encouraging. To provide examples, Faraz said, "I think that you can definitely use calibration in life, like, to know if you are good at your job or you need to improve something," and Zishan commented, "I think this kind of knowledge is general, because with anything in life you have a certain goal and you have, like, a personal expectation. When you start doing something you realise you can do better at it. It's kind of calibration." Additionally, Andreas shared, "Being at work, field, or anywhere, interacting with people, you can be, like, what can I do better in this job, which is calibration, really," and Adnan asserted, "You have to detach yourself from a situation. You have to see a situation from a third person's perspective, and I think that's when metacognition comes in." In another example, Yasir said, "Calibration helps you know what went into knowing what you know and what you actually got. You can see the difference, so you can get better at anything through calibration," while Janani shared, "I found calibration interesting, like, trying to estimate the outcome of any situation."

A Summary of my Overall Findings

In this chapter, I reported my findings as they emerged from the varied layers of my analyses. However, to be explicit, herein I highlight the main findings of my study.

My quantitative analysis suggested an optimistic view in which calibration was perceived to be malleable and could be developed in the context of a classroom intervention, such as mine. The statistical analysis indicated substantial variations in both calibration and bias scores. Although as a group, the students underestimated their task performance outcomes throughout the semester, their collective estimations became more accurate. Notably, as a group, the students made their most accurate estimation on their last self-assessment on their final exam, which suggests that any classroom intervention that aims to improve the calibration accuracy of a group of students in the context of an academic subject, such as mathematics, might take time and explicit training and opportunities for feedback and reflection to be effective. My quantitative analysis further suggested that as a group, students consistently underestimated their task performance throughout the semester as well as on their final exam, and yet, the group's collective calibration accuracy improved over time, which may attest to the complexity and multidimensionality of calibration.

My three-layered blended data analysis provided further insights into the dynamic nature of students' calibration. The findings confirmed the malleability of calibration both conceptually (e.g., as students' metacognitive knowledge developed and they tuned in to improve their own calibration) and in practice (e.g., as students' calibration accuracy improved over time). My blended data analysis suggested that at the start of the intervention a high level of calibration awareness in the form of metacognitive knowledge was not common among the students who also highlighted the influences of personal, behavioural, and environmental factors on their own calibration.

In my view, an important finding of my blended data analysis was students' intentional miscalibration, in which high-achieving students intentionally underestimated their task performance for motivation and self-protection purposes. In my blended data analysis, students' English language facility was an influential factor in the development of their calibration as reported by them. Further, the students indicated mathematics, as a school subject matter, to be influential for their calibration and suggested a bi-directional association between task performance outcomes and calibration accuracy, which is inconsistent with the unidirectional association that is currently suggested in calibration research literature. With regards to students' personal theories of calibration, my blended data analysis indicated that students viewed calibration as a useful concept that was applicable to both academic and non-academic areas of their lives.

In the service of ecological validity, the future research on calibration may benefit from conducting research in actual classrooms with students, giving credence to students' perspectives on what they are confident about, what might influence their calibration, and what kind of support they need (e.g., a metacognitively oriented and inclusive classroom environment, the choice of their own test questions, opportunities for feedback and reflection) in designing classroom interventions to improve students' calibration.

Chapter Four Summary

I started Chapter Four by describing my quantitative data analysis, which explored the numerical and graphical patterns of change in the measures of students' calibration accuracy. Subsequently, I detailed my three rounds of blended data analyses. In the first round, I explored the development of calibration for each individual student by blending the qualitative and quantitative data at 12 testing instances for which the qualitative and quantitative data were simultaneously created by the students. In the second round, I inquired into the development of calibration for the entire group by analysing all the qualitative and quantitative data, except for students' interview data. Finally, in the third round of my analysis, I added the interview data and analysed my entire research data looking for consistencies and inconsistencies in my research findings. I concluded Chapter Four by summarizing the findings of my study and a chapter summary.

In Chapter Five, I discuss my research findings within the context of the calibration research literature as I approach an answer for my research question. Moreover, I comment on the educational implications of this study and its contributions to the theory, teaching practice, and to the advancement of knowledge. I conclude Chapter Five by an outline of the limitations of my findings, my views on the future of calibration research, and my final reflections.

Chapter Five

Discussion

Introduction

Often conducted in controlled experimental settings using diagnostic snapshot designs, calibration research has produced inconclusive or contradictory results in a variety of areas. Questions remain as to whether and how it is possible to develop students' calibration in the context of everyday classrooms, using interventions that utilise monitoring strategy training, curriculum-based classroom tests, students' calibration graphs, and students' reflection notecards. Moreover, there is a disconcerting absence of students' voices, opinions, and self-attributions from the contemporary research on calibration. Through this study, I responded to these concerns within the naturalistic setting of a mathematics classroom in a suburban public high school in central Canada. As such, together with 24 students in a grade 11 mathematics course that I was teaching, we embarked on a journey of learning, reflection, and self-evaluation.

An inspiration for this research was to situate my study within the contemporary scholarly discourse on metacognition research in mathematics education. Further, I aspired to contribute to the body of metacognition studies in mathematics education by presenting calibration as a topic worthy of attention and deep examination due to its role in students' life-long learning and achievement. Central to this study was the exploration of the potential change in mathematics students' calibration accuracy over time and multiple testing sessions. I implemented a mixed methods study, seeking both qualitative and quantitative data. In addition to the quantitative data analysis procedures, I employed three rounds of blended data analyses to analyse and interpret the change in students' calibration in the context of a structured, multifaceted classroom intervention.

In this chapter, I describe how the findings of this study predicate an answer for my research question: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom? Moreover, in this chapter, I convey the educational implications and contributions of my findings, explain the limitations of my findings, provide recommendations for future research, and offer my final reflections.

Approaching an Answer for my Research Question

Johnson and Christensen (2008) contend that mixed methods researchers should employ triangulation to converge the results from the different methods. As described in Chapter Four, my research findings were the result of my analytical synthesis of the quantitative data and the blended data, in the context of the research literature on calibration. The open-ended questionnaires, test performance outcomes, test performance estimations, reflection notecards, calibration graphs, and in-depth interviews were the methods that I used to gain insights into the phenomenon of calibration change. Through comparing and contrasting the results obtained by using these different methods in the context of the vast body of research literature on calibration, I now describe the main points that emerged from my research findings and how these findings are confirmatory and convergent in providing an answer for the research question of my study.

Calibration Is Malleable

Despite the inconclusive results of the experimentally designed studies of calibration as discussed in Chapter Two, the results of this study suggest that calibration is malleable and can be developed. In my study, the calibration accuracy of 18 out of 24 students in the group improved. The change in calibration for four students with initially good calibration accuracy was negligible. However, two students, Rabia and Bilal, became less calibrated over time. This finding was surprising, considering that students had engaged in monitoring training sessions for

14 weeks and were explicitly taught monitoring strategies. The blended data analysis, however, brought into light the influences of the personal (e.g., self-confidence and self-efficacy beliefs), behavioural (e.g., study behaviours), and environmental (e.g., influences of teachers and peers) factors on students' calibration. For instance, as I discussed in Chapter Four, in her comments, Rabia placed an emphasis on her inadequate base knowledge of mathematics and her inability to focus while engaged in task performance, while Bilal shared his intentional underestimation of task performance to avoid disappointment. My quantitative analysis informed me as to *how* calibration changed for Rabia and Bilal (i.e., the measures of discrepancy increased and they became less calibrated over time), but it was the blended data analysis that suggested *why* and *in what way* the change took place. As a researcher, I found the findings of my blended data analysis critical to my understanding of the change in students' calibration.

In a unique case, Zishan consistently *overestimated* his task performance. The qualitative data revealed that this overestimation presumably motivated him. One might consider the possibility that the overall improvement in the measures of calibration accuracy might have been higher than what the quantitative results suggested, had the students not intentionally biased their own estimations. As such, my findings speak to the importance of considering students' views about themselves and what they are confident about in conducting sound calibration research.

Classroom Interventions Have the Potential to Develop Students' Calibration Over Time

Despite the small number of 24 participants, the results of both quantitative and blended data analyses communicated the possible educational significance of this classroom intervention by (a) bringing the importance of self-evaluation of task performance to the students' attention, (b) providing strategies for the monitoring and regulation of task performance while on daily mathematical tasks, and (c) providing the students with a set of metacognitive strategies that are potentially applicable to other areas of life beyond academics. In response to the research question, although the qualitative findings revealed an enhancement in students' calibration awareness and the quantitative results showed an improvement in students' calibration accuracy, the adaptive change in calibration took time and engaged students' personal effort to take effect. This suggests that as a thinking skill, the development of calibration takes time and requires the use of monitoring strategy training and the students' reflection notecards calibration graphs. The convergence of the results of the blended data analysis and the quantitative data analysis suggests that classroom interventions might facilitate the development of students' calibration.

Calibration Is Multidimensional and Complex

The evidence of the recurring notions of students' self-confidence and self-efficacy within my blended data analyses may explain the relatively small improvement in the numerical values of calibration accuracy. Educational researchers have often explored calibration in the light of academics and school-related issues, such as test difficulty, subject matter, and academic expertise. It is possible, however, that the influences on students' calibration are not always academically oriented.

In my study, the blending of the qualitative and quantitative data and the emergent results spoke to the complexity of calibration by revealing that far from being a mere numerical discrepancy, calibration is also a deeply rooted metacognitive and social construct. In response to the research question, I contend that although the results provided optimism by showing that students' calibration could be developed through strategy training, curriculum-based classroom tests, feedback, and reflection, the overall results suggest that the development of calibration is a complex and lengthy process that depends, at least in part, on students' perceptions of themselves and their environment.

Being Unskilled Is not the Only Reason for Miscalibration

The Dunning–Kruger effect (Kruger & Dunning, 1999) postulates that academically unskilled (i.e., low-achieving) students suffer a double misfortune in that they lack both the required academic expertise as well as the skills necessary to be aware of their lacking. In this study, among the 24 participants, the quantitative data showed that Abrar, Andreas, and Shirin performed predominantly at an academically low level throughout the semester. In their comments, they maintained that their calibration neither changed nor improved over time. By blending the qualitative data of these students' comments with the quantitative measures of their calibration, however, I showed that their calibration did change and improve over time. As Kruger and Dunning (1999) and Nietfeld and Schraw (2002) hypothesised, it is possible that the knowledge of the field (e.g., conceptual and procedural content knowledge of mathematics in this study) may be foundational to the accuracy of performance evaluations made in the field. Nevertheless, based on my blended data analysis, ten students (in a group of 24) who intentionally miscalibrated (nine students consistently underestimated, and one student consistently overestimated) were high-achieving students. This group miscalibrated throughout the semester not because they were academically unskilled, but because they were influenced by personal, behavioural, and environmental factors. In responding to the research question, I contend that students' calibration changes in response to their perceptions of the self and the environment.

Stability of Students' Performance Outcomes Explained

As noted in Chapter Four, Figure 4.2 illustrates students mean performance estimations (the dashed line) versus their mean achieved performance outcomes (the full line). In comparison to students' estimations, their performance outcomes were substantially more stable, as the quantitative findings showed that students' performance outcomes were more consistent with less variability than their calibration accuracy measures. This was unprecedented, considering that calibration researchers (e.g., Hacker et al., 2000; Hacker et al., 2008a; Schraw et al., 1993) have noted the stability of students' performance estimations (both predictions and postdictions) to be substantially higher than students' performance outcomes. Further, in contrast to some researchers' (e.g., Roderer & Roebers, 2013) contention that students' calibration scores are highly influenced by test properties, the current findings indicated that students' performance outcomes across two different types of tests (quizzes versus tests and exams) remained considerably stable, but that the calibration measures changed in mean-value and variance across testing episodes. As such, the results suggested that the students possibly took the test properties into account while estimating their performance, which affected their calibration measures. For instance, on a reflection notecard after a quiz, in explaining her miscalibration, Kristian wrote, "I thought it would be easy because it was just a quiz, so I estimated high." Importantly, the students' performance estimations were made *after* answering the test questions, which presumably placed them in a better position for accurate self-estimation of task performance. How students performed mathematically did not seem to be influenced by test properties. Students' lack of confidence in their performance outcomes, however, led to a greater variability in their calibration than in their performance outcomes.

In response to my research question regarding the ways in which calibration might change and develop, although the findings portray an optimistic view in which students' calibration is malleable, they also indicate that calibration change takes time and may require the explicit teaching of metacognitive strategies to take effect. The findings also suggest the possibility that some students may intentionally miscalibrate for personal and motivational reasons. Intentional miscalibration is an important issue that could potentially affect the results of any quantitative study of calibration. This issue can be brought to light only if researchers adopt a mixed methods approach to conduct research on calibration. The findings of my study further indicate that the naturalistic setting used in this study was in fact an important portal for calibration change, and that calibration change depends on not only academic factors, but upon students' perceptions of the self and the world. Finally, in response to the research question, the results of this study reveal that the change in students' calibration was less stable than the change in students' test performance, which also speaks to the malleability of calibration.

I took the long and effortful path of conducting classroom research on calibration to find an answer to my research question: In what ways might students' calibration change and develop in the context of a year 11 mathematics classroom? Before embarking on this journey, being professionally immersed in mathematical thinking for the past three decades, I speculated clear and definitive answers. However, I am now cognizant that through my research I have created more questions than answers. I am also certain that there are no definitive or indisputable answers to my research question.

In theorising about students' calibration change, I have come to realise that the ways in which students try to improve their own calibration are as varied and unique as each individual student. In my study, students' calibration changed not only in personal ways, but also in connection with the social and environmental factors unique to each student. I posit that in order to design classroom interventions that work for students, calibration researchers should consider a *whole person* approach to research, because the percentages that are used to determine calibration are more than mere numbers; they reflect the whole person's views about themselves and the world around them. In my study, those views changed students' calibration and with the change in calibration, some of those views changed, as well, attesting to the complexity and multi-dimensionality of calibration. Hence, it seems appropriate to ask: If calibration is so complex, can students' calibration be developed through the use of classroom interventions? I learned from my study that the answer is yes, however, the intervention should address the whole person rather than to attempt to solicit accurate numbers. For instance, students' confidence about themselves and the subject matter, their proficiency level in the language of instruction, their images of success, and their feelings of achievement and personal expertise are important to address in classroom studies of calibration. By adopting a whole person approach to research, calibration researchers might succeed in conducting classroom interventions that achieve both goals of caring for students, as well as enhancing their calibration.

Educational Implications of This Study

The findings of this study have implications for classroom pedagogy by suggesting that high school mathematics students may benefit from monitoring strategy training incorporated in teachers' daily lessons. To help students develop their calibration, monitoring strategies should be taught explicitly, embedded in relevant mathematical tasks, and discussed, modelled, and reinforced. The findings of this study also posit that metacognitive feedback, performance feedback, and reflection may improve calibration accuracy and encourage self-learning. Yet another implication of the findings is that in teaching mathematics, teachers should be cognizant of students' beliefs about themselves both as an individual and as a mathematics learner, since such beliefs affect students' decision-making and calibration.

What I Learned From my Students

In this study, students contributed to all aspects of my work, possibly without realising how intensely I learned from them and how privileged this knowledge seemed. My unprecedented learning experience went beyond anything that I had learned from my study of the books, research papers, PhD dissertations, journal articles, and reading anything related to calibration that I could find, in a time span that stretched between the year 1900, with Woodworth and Thorndike's paper on judgements of magnitude, and the year 2019, with Stoten's paper on calibration accuracy. Despite benefitting from 120 years of research on calibration, I owe my students a debt of gratitude for teaching me about (a) the inadequacies of language, (b) the undiscovered intricacies of silence, and (c) an appreciation for the immense complexity of calibration.

I call the first dimension of my learning the *inadequacies of language*. In my days of attempting a classroom intervention with students, I became aware that we were all experiencing something for which we did not have the proper language of expression. I came to realise that if calibration researchers were to focus on classroom research, it would be necessary to modify the research methodologies, as well as the language, tone, and modality. This is because sometimes what seems appropriate in theory may be neither useful nor practical in a classroom setting. For instance, in theory, the adaptation of the monitoring strategies by students is expected to take place after one or two brief training sessions. In such cases, while students work individually, they are trained by the researchers who are unknown to them, using the research material that are not of any relevance or motivational value to students. In practice, however, the following prospects need to be considered: (a) Monitoring strategies may need intensive modelling and reinforcement, using familiar material, expressions, and relevant examples; (b) These strategies may require a long period of time, possibly several months, to be internalised and become habitual; and (c) Working in isolation may not be as helpful for the development of students' calibration as working in a classroom setting that embraces social interactions.

The second dimension of my learning was to learn about the *undiscovered intricacies of silence*. I learned to listen more patiently than ever and value the frequent long pause as a remarkable educational tool. I negated the notion of the 'awkward pause' wholeheartedly and learned to appreciate 'silence' as an opportunity, while students struggled to bring their ways of thinking about their thinking out and into the open. This spoke not only to the inadequacies of the language of instruction in dealing with the cognitive phenomena, but also to the fact that talking about ways of thinking seemed to be an unchartered territory for most students.

The third dimension of my learning was *an appreciation for the immense complexity of calibration*, beyond anything that I could have learned from books. Engaging with students in monitoring training sessions, calibration graphing, and calibration sessions week after week was a humbling experience, which illuminated for me the limited reach of the current theories when challenged in the real-world context of a classroom. Although the results of this study supported research findings (e.g., Coutinho, 2007; Roderer & Roebers, 2013) suggesting that classroom interventions may improve students' calibration, my results highlighted new areas of complexity by revealing that to be effective, such interventions need to adopt a whole person approach.

Contributions of This Study

In this section, I detail the contributions of my study in three areas: (a) contributions to the theory, (b) contributions to the educational practice in mathematics education, and (c) contributions to the advancement of knowledge.

Contributions to the Theory

As I noted earlier, since its inception in 1990, Nelson and Narens's theory of metacognition has served as a unifying theory in the study of metacognition and its components. Some calibration researchers (e.g., Boekaerts & Rozendaal, 2010; Efklides, 2011) have criticised the theory for introducing metacognition as a unidimensional construct and for not adequately explaining how students process correct and incorrect confidence judgements, which is essential to the formation of confidence judgements and calibration accuracy. Of considerable importance for the present study was the fact that the theory succinctly explains how improved monitoring skills while on academic tasks may trigger the appropriate control processes. For instance, unless students realise that they are not prepared for an upcoming test (i.e., monitoring processes), they may not decide to review their class notes (i.e., control processes).

The findings of my study suggest that over time, a classroom intervention that involves monitoring strategy training, classroom tests, feedback, and reflection may have an overall positive influence on students' calibration. The results of my blended data analysis also revealed that some students reported a gradual adaptive change in their study behaviour, or recognised the need to do so, which may be related to their developing calibration awareness. This particular outcome which involves a change in students' behaviour provides support for the Nelson and Narens's theory of metacognition by suggesting that when the intervention instructions are delivered at the meta-level (e.g., monitoring strategies; Kuhn, 2016), they may possibly trigger enhanced control processes in students and facilitate behavioural change.

In addition, the findings of this study are aligned with the social cognitive theory (Bandura, 1986) by highlighting the role played by students' self-perceptions in their self-estimations of task performance and their calibration accuracy. Consistent with the social cognitive theory, my study suggests the centrality of metacognitively oriented learning environments to the notion of change in students' metacognition and calibration. The findings provide further support for the social cognitive theory by demonstrating the reciprocity of the influence of calibration accuracy and task performance, and by showing that as a thinking skill, calibration must be explicitly taught, modelled, and reinforced in order to become internalised and habitual.

Contributions to the Teaching Practice in Mathematics Education

The current study was based on a classroom intervention that explored the change in students' monitoring skills (i.e., development of their calibration). Monitoring has been shown to be foundational to human learning (Alexander, 2013; Valdez, 2013) and critical to mathematical problem-solving (Desoete, 2009; Desoete et al., 2003; Stillman & Mevarech, 2010; Roderer & Roebers, 2013; Schoenfeld, 1992). Studies with the specific goal of improving students' calibration accuracy have largely been situated in undergraduate level classrooms (e.g., Bol et al., 2005; Bol & Hacker, 2001; Hacker et al, 2008a; Hadwin & Webster, 2013; Lindsey & Nagel, 2013; Mengelkamp & Bannert, 2010; Nietfeld et al., 2005, 2006a; Nietfeld & Schraw, 2002; Schraw et al., 2014; Schraw et al., 1993; Valdez, 2013) often conducted over a short period of time; no study of calibration that I have been able to locate has explored the development of calibration in the context of a high school mathematics classroom, using a mixed methods methodology.

Moreover, ecological validity (Goodman, 2008) continues to pose a challenge, as calibration researchers often employ experimental laboratory-style designs, producing results that may not always be generalisable to a classroom context. Through this study, I responded to the need for ecologically valid (Goodman, 2008) studies in a mathematics classroom using curriculum-based material.

This study was not strictly pre-planned to elicit a certain power or effect-size but was rather designed to evolve naturally by inviting and including all the students in the course and by using methods embedded in mathematics teaching practice. The design was simple and flexible in order to facilitate adaptation to other courses and grade levels by mathematics teachers. Since I still had to meet the demands and expectations of being a classroom teacher, the elements of the intervention were designed to blend smoothly with the classroom tasks. Although I did not collect any data on students' test anxiety, it is possible that, as I briefly noted in the interview conversations, the students experienced less test anxiety than before knowing that on a test they could choose the questions they thought they knew how to answer. My quantitative analysis showed that over time, students' performance scores rose (see Figure 4.2), suggesting that they possibly became more skilful at choosing the questions that they thought they knew how to solve, indicating better calibration. Mathematics teachers may find the method used in this study a fair method of assessment that provides a model to assess what students think they know, can demonstrate, and are confident about.

In my study, the students demonstrated a gradual improvement in their calibration accuracy, indicative of enhanced monitoring processes. As Valdez (2013) highlighted, monitoring primarily pertains to students' confidence judgements about their knowledge of strategies that support cognition, dictating when and how to execute strategies that may influence their performance. In Valdez's (2013) view, while on an academic task, students' confidence judgements can likely influence how they review, modify, or correct the steps taken towards identifying an answer. Overconfident students, however, may fail to make an effort to detect and correct their mistakes, while underconfident students may spend most of the allocated time on only a few questions and fail to complete the task at hand. The findings of my study were consistent with calibration research (e.g., Butler & Winne, 1995; McCormick, 2003; Schunk, 1996) in suggesting that by incorporating simple calibration training strategies in their daily pedagogy, teachers may bring their students many steps closer to becoming metacognitive self-learners who approach any learning task with efficiency.

Contributions to the Advancement of Knowledge

In this study, I utilised curriculum-based classroom tests and encouraged students' engagement in monitoring strategy training, self-reflection, and feedback to explore the development of students' calibration accuracy. I used two frequently used components of calibration: absolute accuracy and bias. Considering that few studies have been designed with the specific goal of developing students' calibration using classroom interventions (Bol et al., 2005; Hacker et al., 2008a), this study adds to the advancement of the body of knowledge on metacognition and calibration. In addition, the classroom tests and monitoring training methods were innovative insofar as they were designed for mathematical tasks, emphasising calibration accuracy. To date and to my knowledge, no other classroom tests and monitoring training strategies have been offered in calibration research, specific to high school mathematics.

Moreover, my findings add to the depth of prior calibration research findings by proposing that the lack of cognitive or metacognitive strategies may not be the only reason behind students' miscalibration. In particular, my analysis showed that high-achieving students who are often considered high-calibrators may be affected by a curious condition in which they might intentionally miscalibrate. As such, my research findings attest to the complexity and multidimensionality of calibration, proposing that calibration researchers should attend to the reciprocal interactions of personal, behavioural, and environmental influences in describing the elements that may be consequential to students' task estimation and calibration.

In the present study, a reciprocal relationship between calibration accuracy and performance emerged from the students' assertions, adding a new dimension to the unidirectional relationship often cited in calibration research (e.g., Hacker et al, 2008b; Hacker et al., 2000; Jacobson, 1990; Kruger & Dunning, 1999; Maki, 1998) in which students' calibration accuracy depends on their achievement level. In the current study, the students noted that they intended to try to improve their performance scores by improving their calibration accuracy, suggesting that they believed their achievement level depended on improved calibration accuracy, which is also sensible.

More broadly, the findings of this study contributed to the understanding of the nature of students' confidence judgements by providing a first-hand account of the factors that students considered they used to make their confidence judgements—a novel perspective that may be utilised in the design of effective classroom interventions.

Limitations of my Findings

I discussed the limitations of the research design in Chapter Three. In this section, I outline the limitations of the findings of this study. Throughout this classroom intervention, I encountered a number of restrictions that imposed limitations on my findings. For instance, the small number of participants, which was not in my control, limits the generalisability of the results. Further, the nature of the course (i.e., a senior mathematics course) limits the range of the findings. Students' confidence and motivation influenced their calibration accuracy throughout the semester. Expressions of students' self-protection in the form of intentional underestimation superficially inflated the measures of calibration and to varying extents, influenced and biased the quantitative results of my study. Although the classroom studies of calibration such as mine possess ecological validity, I acknowledge that the complexities of a naturally unfolding classroom environment may have affected or limited my research findings.

The qualitative data in the present study revealed a variety of issues of importance that were not addressed in calibration studies. For instance, in my analysis, the students asserted that as a subject matter, mathematics influenced their estimations. It is likely that mathematics, with its specific terminology, symbolism, and structure influenced the students' calibration and limited the findings. This issue was more pronounced for those participants who were learning English as a second or third language, many of whom reported that in their respective home countries, mathematics learning was not so heavily language based as it is within the Canadian mathematics curriculum. It is possible that students' English language facility influenced the current study.

Theoretical limitations were inevitable, considering that the two theories that underpinned this study (i.e., Nelson and Narens's theory and Bandura's social cognitive theory) make certain assumptions, which possibly influenced the findings of this study. To elaborate, the social cognitive theory assumes (Pintrich, 2000; Zimmerman, 2001) that while on learning tasks, students actively construct meaning and establish goals based on various cognitive, social, and environmental influences. This assumption does not take the intentional miscalibration of some students, as seen in this study into account.

Moreover, Nelson and Narens's (1990) theory of metacognition assumes that students' capability to monitor their thinking processes while on learning tasks necessarily leads to their selection of control processes. This view has been challenged by researchers. For instance, Efklides (2008) distinguished between conscious monitoring which is deliberate, and nonconscious monitoring which is ongoing and implicit. In Efklides' (2008) view, ongoing nonconscious monitoring processes do not trigger control processes.

Additionally, even when students monitor diligently, they still have a choice whether to partake in control processes or not. Finally, my qualitative data spoke to the influence of students' self-concepts on their calibration. One way or another, all these considerations may impose limitations on the findings of my study.

Looking Forward

The research on different aspects of cognition, including metacognition, has progressed substantially in the past 4 decades, changing how people view learning, motivation, and achievement. The findings of the current study highlighted the importance and utility of classroom interventions for developing students' monitoring strategies and their calibration. Through future research in this field, scholars may further delineate the nature of students' confidence judgements and the role self-perceptions play in the development of students' calibration.

In this study, in making their confidence judgements, students drew upon a large number of factors that were not always consistent with their identified goals such as improving calibration accuracy. In my attempt to explore the naturally occurring change in students' calibration, I did not provide the students with any guidelines to base their performance estimations on. For instance, students were not asked to make their performance estimations based on their previous performance outcomes, hours of study, or content difficulty. If improving calibration accuracy is intended, however, researchers should be wary of performance estimations made based on unrelated tasks. In the future, calibration researchers may decide to provide clear guidelines to reduce the variability of the bases of students' performance estimations. I propose that such guidelines could lead to reduced variability and more effective classroom interventions.

The findings of the current study illustrated the prominent role of students' perceptions of the self as well as their beliefs about their mathematical capabilities (i.e., their self-efficacy beliefs) on their calibration accuracy—a topic that has rarely been addressed in calibration research. Future research may benefit from giving the due credence to these beliefs and how they affect calibration accuracy. As Anderson, Nashon, and Thomas (2009) elaborated, such perspectives support the ontological and epistemological frameworks that view learning and achievement as occurring holistically and in a broad context.

As I stated in the results section of this dissertation, a noteworthy outcome of the present study emerged when the students reported that they changed their study behaviour and test preparation methods based on what they learned about calibration. This outcome aligns with both Nelson and Narens's (1990) theory and Bandura's (1986) social cognitive theory, suggesting that through the enrichment of students' metacognitive knowledge, classroom interventions such as mine may influence students' learning habits and study behaviour.

Further, the students' views of calibration posit a new topic for future research. It is possible to ask: In what ways might students' personal conceptualisations of calibration influence their actual calibration? This question, to my knowledge, has not been explored in classroom research on calibration. For instance, would the development of calibration take a different path for Rabia, who thinks "calibration means calculating quantities that are unknown" than for Leah, who theorises calibration as "the ability for one to know his or her strengths and weaknesses and to study accordingly"?

In conjunction with Bandura's (1986) social cognitive theory, Nelson and Narens's (1990) theory of metacognition served the present study well. I posit that through the proper synthesis of these two theories, calibration researchers might find a unifying framework through which to explain the dynamic nature of calibration.

Final Reflections

In one of my calibration sessions, I asked the students the following:

- How important is it for a driver to realise that he is driving in the middle of a two-way road?
- How crucial is it for a pilot to know that he is flying too close to the summit or for a surgeon to realise that she has received almost no training for an upcoming surgery?
- How important is it for students to have a clear understanding of their own competency and preparedness before writing an important exam?

The present study rests on the importance of self-evaluation of task performance. I have had the good fortune of combining two of my greatest inspirations—mathematics and education—in my role as a mathematics teacher for the past 3 decades. Although I am grateful, I have often wondered: If I could wish for something that my students would take away with them as they leave the high school behind, what would that be? Would it be a working knowledge of calculus, a good hold on trigonometry, or a thorough understanding of geometry? One may imagine a long list to choose from, but what I truly wish for students is to become a good judge of their own knowledge, understanding, and competencies, because such judgements are the first step toward self-reliance and a successful lifelong journey of learning and achievement.

Conducting educational research on metacognition within the complexities of classroom teaching was effortful but also rewarding. It afforded the participants, including myself, with many instances of pausing to listen to each other and to contemplate on all the unique and personal ways of sense-making in mathematics while engaging in mathematical thinking as well as in thinking about mathematical thinking in the form of monitoring. The intervention served as an opportunity for me to bring the concept of calibration out and into the open and provided the students with a set of thinking strategies with a reach beyond their grade 11 mathematics course.

This doctoral dissertation is the story of my students' calibration as reflected not only in their comments, monitoring judgments, and performance scores, but also in their chatter and sound of laughter, worries and anxieties, and hopes and aspirations. I was the storyteller, but the story was always theirs, not mine. The writing process was also very personal for me; I started to write this dissertation with several quotes from my father in the Summer of 2016, and I lost him in the Fall of 2016. Although my father was known for his love for numbers and precision, I believe he would have been pleased with my attempt to convey not only the numbers and measures, but also the voices, feelings, stories, and memories behind the numbers.

To borrow from Charles Dickens (1987), these are the best of times and the worst of times. With all the comforts of the modern-day life and the advancements of science and technology, after working with teenagers for three decades as a teacher, I have come to admire their brave struggle to endure this age of confusion, isolation, foolishness, and distraction. This thesis was inspired by them and written for them, endeavouring to provide support for their lifelong journey of learning and achievement.

The findings of this classroom study connect students' calibration accuracy to their self-perception and motivation. Unfortunately, despite being an increasingly popular line of inquiry in educational and developmental psychology (Bol & Hacker, 2012), the research on calibration is largely absent from the body of research on metacognition in mathematics education. It is my hope that the current study may attract attention to calibration as a worthy and critical research topic in mathematics education.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? *Learning and Instruction, 24*, 1-3. https://doi.org/10.1016/j.learninstruc.2012.10.003

Alexander, P. A., Dinsmore, D. L., Parkinson, M. M., & Winters, F. I. (2010). Self-regulated learning in academic domains. In B. J. Zimmerman, & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 393-407). Routledge.

Allardice, B. S., & Ginsburg, H. P. (1983). Children's psychological difficulties in mathematics. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 319-349). Academic Press.

Allwood, C. M., Jonsson, A. C., & Granhag, P. A. (2005). The effects of source and type of feedback on child witnesses' metamemory accuracy. *Applied Cognitive Psychology, 19*, 331-344. https://doi.org/10.1002/acp.1071

Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsychology Review*(16), 17-42. https://doi.org/10.1007/s11065-006-9002-x

Anderson, D., Nashon, S. M., & Thomas, G. P. (2009). Evolution of research methods for probing and understanding metacognition. *Research in Science Education, 39*(2), 181-195. https://doi.org/10.1007/s11165-007-9078-1

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*, 3-38. https://doi.org/10.1007/BF01326548

Bandura, A. (1982). The psychology of chance encounters and life paths. *American Psychologist, 37*, 747-755. https://doi.org/10.1037/0003-066X.37.7.747

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W. H. Freeman.

Bandura, A. (2001). Social cognitive theory: An agentic perspetive. *Annual Review of Psychology, 52*, 1-26. https://doi.org/10.1146/annurev.psych.52.1.1

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*, 170-174. https://doi.org/10.1080/00220671.1997.10543773

Bateson, G. (1972). *Steps to an ecology of mind.* Jason Aronson.

Bellon, E., Fias, W., & De Smedt, B. (2019). More than number sense: the additional role of executive functions and metacognition in arithmetic. *Journal of Experimental Child Psychology, 182*, 38-60. https://doi.org/10.1016/j.jecp.2019.01.012

Biryucov, P. (2004). Metacognitive aspects of solving combinatorics problems. *International Journal for Mathematics Teaching and Learning*. 1-19

Bjorkman, M. (1992). Knowledge, calibration, and resolution: A linear model. *Organizational Behavior and Human Decision Processes, 51*, 1-21. https://doi.org/10.1016/0749-5978(92)90002-O

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*, 372-382. https://doi.org/10.1016/j.learninstruc.2009.03.002

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*, 133-151. https://doi.org/10.1080/00220970109600653

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here. *Frontiers in Psychology, 3*, 1-6. https://doi.org/10.3389/fpsyg.2012.00229

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*, 280-287. https://doi.org/10.1016/j.cedpsych.2012.02.004

Bol, L., Hacker, D., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269-290. https://doi.org/10.3200/jexe.73.4.269-290

Bol, L., Riggs, R., Hacker, D. J., Dickerson, D., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*(2), 81-96.

Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics, 57*(2), 213-227. https://doi.org/10.1023/B:EDUC.0000049293.55249.d4

Brown, A. (1980). Metacognitive development and reading. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 453-481). Erlbaum.

Brown, A. L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 367-406). Halsted Press.

Brown, A. L. (1994). The advancement of learning. *Educational Researcher, 23*(8), 4-12.

Bryman, A. (2001). *Social research methods.* Oxford University Press.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3-14. https://doi.org/10.3102/0013189X032009003

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245-281. https://doi.org/10.2307/1170684

Cai, J., & Lester, F. (2010). Why is teaching with problem solving important to student learning? In J. Quander (Ed.), *Problem solving research brief* (pp. 1-5). National Council of Teachers of Mathematics.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition Learning, 11*, 215-235. https://doi.org/10.1007/s11409-015-9142-6

Campione, J., Brown, A., & Connell, M. (1989). Metacognition: On the importance of knowing what you are doing. In R. Charles, & E. Silver (Eds.), *The teaching and assessing of mathematics problem solving* (pp. 93-114). National Council of Teachers of Mathematics.

Cardelle-Elawar, M. (1992). Effects of teaching metacognitive skills to students with low mathematics ability. *Teaching and Teacher Education, 8*, 109-121. https://doi.org/10.1016/0742-051x(92)90002-k

Cardelle-Elawar, M. (1995). Effects of metacognitive instructions on low-achievers in mathematics problems. *Teaching and Teacher Education, 11*(1), 81-95. https://doi.org/10.1016/0742-051X(94)00019-3

Carr, M., & Jessup, D. L. (1995). Cognitive and metacognitive predictors of mathematics strategy use. *Learning and Instruction, 7*, 235-247. https://doi.org/10.1016/1041-6080(95)90012-8

Carr, M., Alexander, J., & Folds-Bennett, T. (1994). Metacognition and mathematics strategy use. *Applied Cognitive Psychology, 8*, 583-595. https://doi.org/10.1002/acp.2350080605

Chen, P. P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*, 77-90. https://doi.org/10.1016/j.lindiff.2003.08.003

Chen, P. P., & Zimmerman, B. J. (2007). A cross-national comparison study on the accuracy and self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education, 75*, 221-244. https://doi.org/10.3200/jexe.75.3.221-244

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Lawrence Earlbaum and Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cohors-Fresenborg, E., & Kaune, C. (2001). Mechanisms of the taking effect of metacognition in understanding processes in mathematics teaching. In G. Torner, R. Bruder, N. Neill, A. Peter-Koop, & B. Wollring (Eds.), *Developments in mathematics education in German-speaking countries: Selected papers from the annual conference on didactics of mathematics* (pp. 29-38). Franzbecker.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* Oxford University Press.

Cornoldi, C., Lucangeli, D., Caponi, B., Falco, G., Focciatti, R., & Todeschini, M. (1995). *Mathematica e metacognizione.* Erickson.

Coutinho, S. A. (2007). The relationship between goals, metacognition, and academic success. *Educate, 7*, 39-47.

Crane, N., Zusho, A., Ding, Y., & Cancelli, A. (2017). Domain-soecific metacognitive calibration in children with learning disabilities. *Contemporary Educational Psychology*(50), 72-79. https://doi.org/10.1016/j.cedpsych.2016.09.006

Crissman, J. K. (2006). *The design and utilization of effective worked examples: A meta-analysis.* Unpublished doctoral dissertation, The University of Nebraska-Lincoln.

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology, 80*(2), 131-142. https:// doi.org/10.1037/0022-0663.80.2.131

Cunha, F., & Heckman, J. (2006). Investing in our young people. *Rivista Internazionale di Scienze Sociali, 117*(3/4), 387-417.

de Bruin, A. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: from cognitive psychology to the classroom. *Learning and Instruction, 22*, 245-252. https://doi.org/10.1016/j.learninstruc.2012.01.003

Dembo, M. H., & Jakubowski, T. G. (2003). The influence of self-protective perceptions on the accuracy of test prediction. *Paper presented at the annual meeting of the American Educational Research Association.*

Denscombe, M. (2010). *The good research guide for small-scale social research participants* (4th ed.). Open University Press.

Denzin, N. K., & Lincoln, Y. S. (2005). The discipline and practice of qualitative research. In N. K. Denzin, & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (3 ed., pp. 1-32). Sage.

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning*, 189-206. https://doi.org/10.1007/s11409-008-9026-0

Desoete, A. (2009). Metacognitive prediction and evaluation skills and mathematical learning in third-grade students. *Educational Research and Evaluation: An International Journal of Theory and Practice, 15*(5), 435-446. https://doi.org/10.1080/13803610903444485

Desoete, A., & Roeyers, H. (2006). Metacognitive macro-evaluations in mathematical problem-solving. *Learning and Instruction, 16*, 12-25. https://doi.org/10.1016/j.learninstruc.2005.12.003

Desoete, A., & Veenman, M. V. (2006). Metacognition in mathematics: Critical issues on nature, theory, assessment and treatment. In A. Desoete, & M. V. Veenman (Eds.), *Metacognition in mathematics education* (pp. 1-10). Nova Science Publishers.

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can off-line metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*(1), 188-200. https://doi.org/10.1037/0022-0663.95.1.188

Dickens, C. (1987). *A Tale of Two Cities.* (H. Bloom, Ed.) Chelsea House Publishers.

DiFrancesca, D., Nietfeld, J., & Cao, L. (2015). A comparison of high and low achieving students on self-regulated learning variables. *Learning and Individual Differences, 45*, 228-236. https://doi.org/10.1016/j.lindif.2015.11.016

Dignath, C., & Buttner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*, 231-264. https://doi.org/10.1007/s11409-008-9029-x

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4-14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*, 391-409. https://doi.org/10.1007/s10648-008-9083-6

Dublinsky, E. (1991). Reflective abstraction in advanced mathematical thinking. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 95-126). Springer. https://doi.org/10.1007/0-306-47203-1

Dunlosky, J., & Lipko, A. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228-232. https://10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* SAGE.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*, 58-61. https://doi.org/10.1016/j.learninstruc.2012.05.002

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychology Review, 111*, 524-542. https://doi.org/10.1037/0003-295x.111.2.524

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Sciences, 12*(3), 83-87. https://doi.org/10.1111/1467-8721.01235

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277-287. https://doi.org/10.1027/1016-9040.13.4.277

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL Model. *Educational Psychologist, 46*, 6-25. https://doi.org/10.1080/00461520.2011.538645

Ericsson, K., & Simon, H. A. (1980). Verbal reports as data. *Psychology Review*(87), 215-251.

Ferrari, M., Robinson, D., & Yasnitsky, A. (2010). Wundt, Vygotsky, and Bandura: A cultural-historical science of consciousness in three acts. *History of Human Sciences, 23*(3), 95-118. https://doi.org/10.1177/0952695110363643

Field, A. (2009). *Discovering statistics using IBM SPSS.* SAGE.

Finn, B., & Tauber, S. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about process fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27*, 567-586. https://doi.org/10.1007/s10648-015-9313-7

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: the approrriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552-564. https://doi.org/10.1037/0096-1523.3.4.552

Flannelly, L. T. (2001). Using feedback to reduce students' judgment bias on test questions. *Journal of Nursing Education, 40*, 10-16.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget.* Van Nostrand.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. Resnick (Ed.), *The nature of intelligence* (pp. 231-235). NJ: Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new era of cognitive development inquiry. *American Psychologist, 34*, 906-911. https://doi.org/10.1037/0003-006X.34.10.906

Flavell, J. H. (1985). *Cognitive development* (2 ed.). Prentice-Hall.

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail, & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3-30). Erlbaum Associates.

Fosnot, C. T. (Ed.). (1996). *Constructivism: Theory, perspectives, and practice.* Teachers College Press.

Fox, E., & Riconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review, 20*(4), 373-389. https://doi.org/10.1007/s10648-008-9079-2

France-Kaatrude, A., & Smith, W. P. (1985). Social comparison, task motivation, and the development of self-evaluative standards in children. *Developmental Psychology, 21*, 1080-1089. https://doi.org/10.1037/0012-1649.21.6.1080

Garcia, T., Rodriguez, C., Gonzalez-Castro, P., Gonzalez-Pienda, J., & Torrance, M. (2016). Elementary students metacognitive processes and post performance calibration on mathematical problem-solving tasks. *Metacognition Learning, 11*, 139-170. https://doi.org/10.1007/s11409-015-9139-1

Garner, R., & Alexander, P. A. (1989). Metacognition: Answered and unanswered questions. *Educational Psychologist, 24*, 143-158. https://doi.org/10.1207/s15326985ep2402_2

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16*(3), 163-176. http://www.jstor.org/stable/748391

Gergen, K. J. (1985). The social constructivist movement in modern psychology. *American Psychologist, 40*(3), 175-266. https://doi.org/10.1037/0003-066X.40.3.266

Glaser, R., & Chi, M. T. (1988). Overview. In M. T. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise.* Lawrence Erlbaum Associates.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension . *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702-718. https://doi.org/10.1037/0278-7393.11.1-4.702

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory and Cognition, 15*(1), 84-93. https://doi.org/10.3758/BF03197714

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119-136. https://doi.org/10.1037/0096-3445.116.2.119

Goodman, S. (2008). The generalizability of discursive research. *Qualitative Research in Psychology, 5*, 265-275. https://doi.org/10.1080/14780880802465890

Goos, M. (1994). Metacognitive decision making and social interactions during paired problem solving. *Mathematics Education Research Journal, 6*(2), 144-165.

Goos, M. (2002). Understanding metacognitive failure. *Journal of Mathematical Behavior, 21*, 283-302. https://doi.org/10.1016/s0732-3123(02)00130-x

Gravalia, L. S., & Gredler, M. E. (2002). An exploratory study of academic goal setting, achievement, calibration, and self-regulated learning. *Learning Journal of Instructional Psychology, 29*, 221-230.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation design. *Educational Evaluation and Policy, 11*, 255-274. https://doi.org/10.2307/1163620

Guba, E. G. (1987). What have we learned about naturalistic evaluation? *Evaluation Practice, 8*, 23-43. https://doi.org/10.1016/S0886-1633(87)80037-5

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin, & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105-117). Sage.

Guion, L. A., Diehl, D. C., & McDonald, D. (2002). *Triangulation: Establishing the validity of qualitative studies.* University of Florida.

Gutierrez de Blume, A. P. (2017). The effects of strategy training and an extrinsic incentive on fourth- and fifth-grade students' performance, confidence, and calibration accuracy. *Cogent Education, 4*, 1-17. https://doi.org/10.1080/2331186x.2017.1314652

Gutierrez de Blume, A., Wells, P., Davis, A., & Parker, J. (2017). You can sort of feel it: exploring metacognition and the feeling of knowing among undergraduate students. *The Qualitative Report, 22*(7).

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model for metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1-10. https://doi.org/10.1016/j.learninstruc.2016.02.006

Hacker, D. J., Bol, L., & Bahbahani, K. (2008a). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition Learning, 3*, 101-121. https://doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., & Keener, M. C. (2008b). Metacognition in education: Focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429-455). Psychology Press.

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160-170. https://doi.org/10.1037/0022-0663.92.1.160

Hacker, D. J., Keener, M. C., & Kircher, J. C. (2009). Writing is applied metacognition. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 154-170). Taylor & Francis.

Hacker, D., Dunlosky, J., & Graesser, A. C. (Eds.). (2009). *Handbook of metacognition in education.* Routledge.

Hadwin, A. F., & Webster, E. (2013). Calibration in goal setting: Examining the nature of judgments of confidence . *Learning and Instruction, 24*, 37-47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Hart, L. C., Smith, S. Z., Swars, S. L., & Smith, M. E. (2009). An examination of research methods in mathematics education. *Journal of Mixed Methods Research, 3*(1), 26-41. https://doi.org/10.1177/1558689808325771

Hartwig, M. K., Was, C. A., Isaacson, R. M., & Dunlosky, J. (2012). General knowledge monitoring as a predictor of in-class exam performance. *British Journal of Educational Psychology, 82*(3), 456-468. https://doi.org/10.1111/j.2044-8279.2011.2038x

Hartwig, M., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80-90.

Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction, 24*, 62-66. https://doi.org/10.1016/j.learninstruc.2012.05.009

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Hennessey, M. G. (1999). Probing the dimensions of metacognition: Implications for conceptual change teaching-learning. Paper presented at the annual meeting of the National Association for Research in Science Teaching.

Higham, P. A. (2013). Regulating accuracy on university tests with the plurality option. *Learning and Instruction, 24*, 26-36. https://doi.org/10.1016/j.learninstruc.2012.08.001

Horgan, D. (1990). Competition, calibration, and motivation. *Teaching, Thinking, and Problem Solving, 12*, 5-10.

Hostetler, K. (2005). What is "good" educational research? *Educational Researcher, 34*(6), 17-21. https://doi.org/10.3102/0013189X034006016

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition Learning, 4*, 161-176. https://doi.org/10.1007/s11409-009-9042-8

Isaacson, R. M., & Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning: Academic success and reflection on learning. *Journal of the Scholarship of Teaching and Learning, 6*(1), 39-55.

Jacobs, J., & Paris, S. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22*, 255-278. https://doi.org/10.1207/s15326985ep220384_4

Jacobse, A. E., & Harskamp, E. G. (2012). Toward efficient measurement of metacognition in mathematics problem solving. *Metacognition Learning, 7*, 133-149.

Jacobson, R. (1990). Unobservable effects and business performance . *Marketing Science, 9*(1), 74-85.

Jamieson-Noel, D. L., & Winne, P. H. (2003). Comparing self-reports to traces of studying behavior as representations of students studying and achievement . *German Journal of Educational Psychology, 17*, 159-171.

Johnson, A. C., & Allwood, C. M. (2003). Stability and variability in realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*, 559-574. https://doi.org/10.1016/S0191-8869(02)00028-4

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches* (Third ed.). Sage.

Johnson, B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, S. M., & White, G. (1971). Self-observation as an agent of behavioral change. *Behavior Therapy, 2*, 488-492.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: heuristics and biases.* Cambridge University Press.

Karpov, Y. V. (2014). *Vygotsky for educators.* Cambridge University Press.

Kaune, C. (2006). Reflection and metacognition in mathematics education: Tools for the improvement of teaching quality. *Zentralblatt fur Didaktik der Mathematik, 38*(4), 350-360. https://doi.org/10.1007/BF02652795

Keren, G. (1991). Calibration and probability judgment: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273. https://doi.org/10.1016/0001-6918(91)90036-y

Kirschner, P. A., & van Merrienboer, J. J. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist, 48*, 169-183. https://doi.org/10.1080/00461520.2013.804395

Klassen, R. M. (2002). A question of calibration: A review of self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly, 25*, 88-102. https://doi.org/10.2307/1511276

Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the performance of business students on examinations . *Journal of Education for Business, 79*(4), 217-224. https://doi.org/10.3200/JOEB.79.4.217-224

Koriat. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge University Press.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning . *Journal of Experimental Psychology: General, 126*, 297-316.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General , 140*, 117-139. https://doi.org/10.1037/a0022171

Koriat, A., & Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder . *Consciousness and Cognition, 9*, 193-202. https://doi.org/10.1006/ccog.2000.0436

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147-162.

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training . *American Educational Research Journal, 40*(1), 281-310. https://doi.org/10.3102/00028312040001281

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessment. *Journal of Personality and Social Psychology, 77*(6), 1121-1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kruger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180-188.

Kuhn, D. (2016). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178-181. https://doi.org/10.1111/1467-8721.00088

Kuhn, D., & Dean, D. (2010). A bridge between cognitive psychology and educational practice. *Theory into Practice, 43*(4), 268-273. https://doi.org/10.1207/s15430421tip4304_4

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative results. *Metacognition Learning, 5*, 173-194. https://doi.org/10.1007/s11409-010-9056-2

Lai, E. R. (2011). *Motivation: A literature review.* Pearson Research Report.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-test and ANOVAs. *Frontiers in Psychology, 4*, 1-12. https://doi.org/10.3389/fpsyg.2013.00863

Legg, A. M., & Locker, L. J. (2009). Math performance and its relationship to math anxiety and metacognition. *North American Journal of Psychology, 11*, 471-486.

Lesh, R. (1982). *Metacognition in mathematical problem solving: Unpublished Manuscript.* Northwestern University School of Education.

Lester, F. K. (1982). Issues in teaching mathematical problem solving in elementary grades. *School Science & Mathematics, 82*(2), 93-98.

Lester, F. K., & Garofalo, J. (Eds.). (1982). *Mathematical problem solving: Issues in research.* Franklin Institute.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slavic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge University Press.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345-391. https://doi.org/10.1006/ceps.1998.0972

Lincoln, Y., & Guba, E. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. In D. D. Williams (Ed.), *Naturalistic Evaluation* (pp. 73-84). Jossey-Bass.

Lindsay, B. A., & Nagel, M. I. (2015). Do students know what they know? Exploring the accuracy of students' self-assessments. *Physical Education Review: Special Issue, 11*(1). https://doi.org/020103-1-020103-11

Lindsey, B. A., & Nagel, M. (2013). Assessing students' metacognitive calibration with knowledge surveys. *AIP Conference Proceedings*, (pp. 258-261). https://10.1063/1.4789701

Lucangeli, D., Cornoldi, C., & Tellarini, M. (1998). Metacognition and learning disbilities in mathematics. In T. E. Scruggs, & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disabilities* (pp. 219-244). JAI Press.

Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findingd in text expectancy generalize to classroom outcomes? *Review of Educational Research, 61*, 94-106.

Maggioni, L., & Parkinson, M. (2008). The role of teacher epistomic cognition, epistomic beliefs, and calibration in instruction. *Educational Psychology Review, 22*(4), 445-461.

Maki, R. H. (1998). Test predictions over test material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in theory and practice* (pp. 117-144). Erlbaum.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*(4), 723-731. https://doi.org/10.1037/0022-0663.97.4.723

Martinez, M. E. (2006). What is metacognition. *Phi Delta Kappan*, 696-699.

McCormick, C. B. (2003). Metacognition and learning. In W. M. Reynolds, & G. E. Miller, *Handbook of psychology: Educational psychology* (pp. 77-102). Wiley.

Medina, J. (2008). *Brain Rules.* Pear Press.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its consequences for young adolescents' course enrollment intentions and performances in mathematics . *Journal of Educational Psychology, 82*, 60-70.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory and Cognition, 38*(4), 441-451. https://doi.org/10.3758/MC.38.4.441

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174-179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* Massachusetts Institute of Technology.

Miller, A. L. (2012). Investigating social desirability bias in student self-report surveys. *Educational Research Quarterly, 36*(1), 30-47.

Miller, H., & Bichsel, J. (2004). Anxiety, working memory, gender, and math performance. *Personality and Individual Differences, 37*(3), 591-600.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam prediction. *Metacognition Learning, 6*, 303-314. https://doi.org/10.1007/s11409-011-9083-7

Montague, M. (1998). Research on metacognition in special education. In T. E. Scruggs, & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disabilities* (pp. 151-183). Elsevier.

Morgan, M. (1987). Self-monitoring and goal-setting in private study. *Contemporary Educational Psychology, 12*, 1-6. https://doi.org/10.1016/S0361-476X(87)80033-4

Moses, R. P., & Cobb, C. E. (2001). *Radical equations: Math literacy and civil rights.* Beacon Press.

Muis, K. R., Winne, P. H., & Ranellucci, J. (2016). The role of calibration bias and performance feedback in achievement goal regulation. *International Education Research, 4*(1), 14-36. https://doi.org/10.12735/ier.v4ilp14

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

National Council of Teachers of Mathematics (NCTM). (2007). *Mathematics teaching today: Improving practice, improving student learning.*

Nelson, T. O. (1996). Gamma is a measure of accuracy predicting performance on one item relative to another item, not the absolute performance on an individual item: comments on Schraw. *Applied Cognitive Psychology, 10*, 257-260.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267-270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125-173). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). MIT Press.

Nelson, T. O., Kruglanski, A. W., & Jost, J. T. (1998). Knowing thyself and others: Progress in metacognitive social psychology. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 69-89). SAGE.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *Journal of Educational Research, 95*, 131-142. https://doi.org/10.1080/00220670209596583

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education, 74*, 7-28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006a). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy and self-efficacy. *Metacognition Learning, 1*, 159-179. https://doi.org/10.1007/s10409-006-9595-6

Nietfeld, J., Enders, C., & Schraw, G. (2006b). A Monte Carlo comparison of measures of relative and absolute accuracy. *Educational and psychological measurement, 66*(2), 258-271. https://doi.org/10.1177/0013164404273945

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes . *Psychological Review, 84*(3), 231-259.

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2 ed.). McGraw Hill.

O'Connor, M. (1989). Models of human behavior and confidence in judgment: A review. *Internatinal Journal of Forecasting, 5*, 159-169. https://doi.org/10.1016/0169-2070(89)90083-6

Oerke, B., & Bogner, F. X. (2013). Social desirability, environmental attitudes, and general ecological behaviour in children. *International Journal of Science Education, 35*(5), 713-733.

Ontario Ministry of Education. (2010). *Growing success: assessment, evaluation, and reporting in Ontario's schools: covering grades 1-12.* Ministry of Education.

Onwuegbuzie, A. J., & Leech, N. L. (2006). Linking research questions to mixed methods data analysis procedures 1. *The Qualitative Report, 11*(3), 474-498.

Onwuegbuzie, A. J., Johnson, R. B., & Collins, K. M. (2009). A call for mixed analysis: A philosophical framework for combining qualitative and quantitative. *International Journal of Multiple Research Approaches, 3*, 114-139.

Onwuegbuzie, J. A., & Combs, J. P. (2010). Emergent data analysis techniques in mixed methods research: A synthesis. In A. Tashakkori, & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social and Behavioral Research* (pp. 397-430). SAGE Publications. https://doi.org/10.41359781506335193

Ormrod, J. E. (1999). *Human learning* (3rd ed.). Merrill.

Ots, A. (2013). Third graders' performance predictions: calibration deflections and academic success. *European Journal of Psychology of Education, 28*(2), 223-237.

Ozsoy, G. (2012). Investigation of fifth grade students' mathematical calibration skills. *Educational Sciences: Theory & Practice, 12*(2), 1190-1194.

Ozturk, N. (2017). Assessing metacognition: theory and practice. *International Journal of Assessment Tools in Education, 4*(2), 134-148. https://doi.org/10.21449/ijate.298299

Pajares, F. (2002). *Overview of social cognitive theory and self-efficacy.* Emory College.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 20*, 426-443.

Pajares, F., & Schunk, D. H. (2002). Self and self-belief in psychology and education: An historical perspective. In J. Aronson (Ed.), *Improving academic achievement* (pp. 5-25). Academic Press.

Palincsar, A. S. (1986). The role of dialogue in providing scaffolded instruction. *Educational Psychologist, 21*, 73-98. https://doi.org/10.1080/00461520.1986.9653025

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117-175.

Palmer, J. P. (2007). *The courage to teach.* John Wiley & Sons.

Papaleontiou-Louca, E. (2003). The concept and instruction of metacognition. *Teacher Development, 7*, 9-30. https://doi.org/10.1080/13664530300200184

Paris, S. G., & Oka, E. R. (1989). Strategies for comprehending text and coping with reading difficulties. *Learning Disability Quarterly, 12*(1), 32-42. https://doi.org/10.2307/1510250

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones, & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15-52). Erlbaum.

Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: A program to improve children's awareness and comprehension. *Journal of Educational Psychology, 76*, 1239-1252. https://doi.org/10.1037/0022-0663.76.6.1239

Parkinson, M. M., Dinsmore, D. L., & Alexander, P. A. (2010). Calibration calibration: Towards conceptual clarity and agreement in calculation. *Paper presented at the annual meeting of the American Educational Research Association.*

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles concepts and evidence. *Psychological Science in the Public Interest, 9*, 105-119.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2 ed.). Sage Publications.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Sage.

Pennequin, V., Sorel, O., Nanty, I., & Fontaine, R. (2010). Metacognition and low achievement in mathematics: The effect of training in the use of metacognitive skills to solve mathematical problems. *Thinking & Reasoning, 16*(3), 198-220. https://doi.org/10.1080/13546783.2010.509052

Persky, A., & Dinsmore, D. (2019). Metacognitive changes and sources of confidence judgments in health professions classroom learning. *Currents in Pharmacy Teaching and Learning, 11*, 338-345. https://doi.org/10.1016/j.cptl.2019.01.005

Piaget, J. (1970). *Genetic epistemology.* Columbia University Press.

Pieschl, S. (2009). Metacognitive calibration-an extended conceptualization and potential applications. *Metacognition Learning, 4*, 3-31. https://doi.org/10.1007/s11409-008-9030-4

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). Academic.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385-407. https://doi.org/10.1007/s10648-004-0006-x

Pintrich, P. R., & Schunk, D. K. (2002). *Motivation in education: Theory, research, and applications* (2nd. ed.). Merrill/Prentice Hall.

Pintrich, P. R., Wolters, C., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw, & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43-97). Buros Institute of Mental Measurements.

Polya, G. (1957). *How to solve it.* Princeton University Press.

Polya, G. (1973). *How to solve it: A new aspect of mathematical method.* Princeton University Press.

Pressley, M. (1986). The relevance of good strategy user model to the teaching of mathematics. *Educational Psychologist, 21*, 139-161. https://doi.org/10.1080/00461520.1986.9653028

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Lawrence Erlbaum Associates.

Pressley, M., Borkowski, J. G., & Schneider, W. (1989). Good information processing: What it is and what education can do to promote it. *International Journal of Educational Research, 13*, 857-867. https://doi.org/10.1016/0883-0355(89)90069-4

Pressley, M., Ghatala, E., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main idea and don't realize it: Confidence in short-answer and multiple-choice comprehension questions. *Reading Research Quarterly*, 232-249.

Puncochar, J. M., & Fox, P. W. (2004). Confidence in individual and group decision making: When "two heads are worse than one". *Journal of Educational Psychology, 96*, 582-591.

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Jounal of Advanced Academics, 20*(1), 18-41. https://doi.org/10.4219/jaa-2008-869

Reid, A., Morrison, G. R., & Bol, L. (2017). Knowing what you know: Improving metacomprehension and calibration accuracy in digital text. *Educational Technology Research and Development , 65*, 29-45. https://doi.org/10.1007/s11423-016-9454-5

Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology*, 23-32. https://doi.org/10.1002/1099-0720(200101/02)15:1<23::AID-ACP681>3.0.CO2-J

Rinne, L. F., & Mazzocco, M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS ONE, 9*(7). https://doi.org/10.1371/journal.pone.0098663

Roderer, T., & Roebers, C. (2013). Children's performance estimation in mathematics and science tests over a school year: A pilot study. *Electronic Journal of Research in Educational Psychology, 11*(1), 5-24. https://doi.org/10.7892/boris.45238

Roeser, R. W., Eccles, J. S., & Strobel, K. R. (1998). Linking the study of schooling and mental health: Selected issues and empirical illustrations at the level of the individual. *Educational Psychologist, 33*, 153-176. https://doi.org/10.1207/s15326985ep3304_2

Rutherford, T. (2017a). The measurement of calibration in real context. *Learning and Instruction, 47*, 33-42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Rutherford, T. (2017b). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226-237. https://doi.org/10.1016/j.cedpsych.2017.03.001

Sale, J., Lohfeld, L., & Brazil, K. (2002). Revisiting the qualitative-quantitative debate: Implications for mixed-methods research. *Quality and Quantity, 36*, 43-53.

Salter, P. (2012). Developing self-regulated learners in secondary schools. *Joint AARE APERA International Conference*, (pp. 1-11).

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: major trends and implications for education. *Mind, Brain, and Education, 2*(3), 114-121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *The International Journal on Mathematics Education, 42*, 149-161.

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky, & B. Bjork (Eds.), *A handbook of memory and metacognition* (pp. 391-409). Erlbaum.

Schoenfeld, A. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334-370). MacMillan.

Schoenfeld, A. H. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education, 10*(3), 173-187.

Schoenfeld, A. H. (1981). Episodes and executive decisions in mathematical problem solving. In R. L. Landau (Ed.), *Acquisition of mathematics concepts and processes* (pp. 345-395). Academic Press.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* Academic Press.

Schoenfeld, A. H. (1987a). Polya, problem solving, and education. *Mathematics Magazine, 60*(5), 283-291. https://doi.org/10.1080/0025570x.1987.11977325

Schoenfeld, A. H. (1987b). What is all the fuss about metacognition? In A. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 334-370). Erlbaum.

Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context . *Issues in Education, 4*(1), 1-94. https://doi.org/10.1016/s1080-9724(99)80076-7

Schoenfeld, A. H. (1999). Looking toward the 21st century: Challenges of educational theory and practice. *Educational Researcher, 28*(7), 4-14. http://www.jstor.org/stable/1176136

Schoenfeld, A. H. (2000). Purposes and methods of research in mathematics education. *Notices of the AMS, 47*(6), 641-649.

Schoenfeld, A. H. (2002). Looking for leverage: Issues of classroom research on "Algebra for All". *Proceedings of the International Conference of the Teaching of Mathematics.*

Schoenfeld, A. H. (2005). On learning environments that foster subject-matter competence. In L. Verschaffel, E. De Corte, G. Kanselaar, & M. Valcke (Eds.), *Powerful environments for promoting deep conceptual and strategic learning* (pp. 29-44). Studia Paedagogica.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the what works clearing-house to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13-21.

Schoenfeld, A. H. (2009a). Instructional research and the improvement of practice. In J. D. Bransford, D. J. Stipek, N. J. Vye, L. M. Gomez, & D. Lam (Eds.), *The role of research in educational improvement* (pp. 161-188). Harvard Education Press.

Schoenfeld, A. H. (2009b). Why do we teach? *Kappa Delta Pi Record, 46*(1), 26-29.

Schoenfeld, A. H. (2010). Reflections of an accidental theorist. *Journal for Research in Mathematics Education, 41*(2), 104-116.

Schoenfeld, A. H. (2011). Toward professional development for teachers grounded in a theory of decision-making. *ZDM Mathematics Education, 43*, 457-469. https://doi.org/10.1007/s11858-011-0307-8

Schoenfeld, A. H. (March 1982). *Beyond the purely cognitive: Metacognition and social cognition as driving forces in intellectual performance.* Paper presented at the annual American Educational Research Association meeting.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19*, 143-154.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9*, 321-332. https://doi.org/10.1002/acp.2350090405

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgment. *The Journal of Experimental Education, 65*(2), 135-146. https://doi.org/10.1080/00220973.1997.9943788

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*(1-2), 113-125. https://doi.org/10.1023/A:1003044231033

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition Learning*(4), 33-45. https://doi.org/10.1007/s11409-008-9031-3

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415-427). Taylor & Francis.

Schraw, G., & DeBacker Roedel, T. (1994). Test difficulty and judgment bias. *Memory and Cognition, 22*(1), 63-69. https://doi.org/10.3758/bf03202762

Schraw, G., & Dennison, R. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460-475. https://doi.org/10.1006/ceps.1994.1033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-374. https://doi.org/10.1007/BF02212307

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1-2), 111-139. https://doi.org/10.1007/s11165-005-3917-8

Schraw, G., Kuch, F., & Gutierrez, A. P. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology, 999*. https://doi.org/10.1037/a0036653

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constrains on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Schunk, D. (1996). Goal and self-evaluative influences during children's metacognitive skill learning. *American Educational Research Journal, 33*(2), 359-382. https://doi.org/10.3102/00028312033002359

Schunk, D. H. (1984). Self-efficacy and classroom learning. *Psychology in the Schools, 22*, 208-223.

Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review, 1*, 173-208. https://doi.org/10.1007/bf01320134

Schunk, D. H. (1989). Social cognitive theory and self-regulated learning. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice* (pp. 83-110). Springer Verlag.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review, 20*, 463-467. https://doi.org/10.1007/s10648-008-9086-3

Schunk, D. H. (2020). *Learning Theories* (8th ed.). Pearson. https://10.1037/13273-005. https://www.pearsonhighered.com/assets/preface/0/1/3/4/0134893751.pdf

Shapiro, E. S. (1984). Self-monitoring procedures. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavior assessment: Principles and procedures* (pp. 148-165). Pergamon.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments . *Organizational Behavior and Human Decision Processes, 42*, 271-278. https://doi.org/10.1016/0749-5978(88)90001-5

Sheldrake, R. (2016). Confidence as motivational expression of interest, utility, and other influences: exploring under-confidence and over-confidence in science students at secondary school. *International Journal of Educational Research, 76*, 50-65. https://doi.org/10.1016/j.ijer.2015.12.001

Sheldrake, R., Mujtaba, T., & Reiss, M. J. (2014). Calibration of self-evaluations of mathematical ability for students in England aged 13 and 15, and their intentions to study non-compulsory mathematics after age 16. *International Journal of Educational Research, 64*, 49-61. https://doi.org/10.1016/j.ijer.2013.10.008

Silver, E. (2004). Ella Minnow Pea: an allegory of our times? *Journal for Research in Mathematics Education, 35*(3), 154-156.

Silver, E. A. (1982a). Knowledge organization and mathematical problem solving. In F. K. Garofalo, *Mathematical problem solving: Issues in research* (pp. 15-25). Franklin Institute Press.

Silver, E. A. (1982b). *Thinking about problem solving: Toward an understanding of metacognitive aspects of problem solving.* Unpublished manuscript, San Diego State University.

Simon, M. K. (2011). *Dissertation and scholarly research: Recipes for success* (11 ed.). Dissertation Success, LLC.

Skemp, R. R. (1979). Goals of learning and qualities of understanding. *Mathematics Teaching, 88*, 44-49.

Stacey, K. (1991). Making optimal use of mathematical knowledge. *Australian Journal of Remedial Education, 22*(4), 6-10.

Steffe, L. P., & Gale, J. (1995). *Constructivism in education.* Lawrence Erlbaum.

Stillman, G., & Mevarech, Z. (2010). Metacognition research in mathematics education: From hot topic to mature field. *ZDM Mathematics Education, 42*, 145-148. https://doi.org/10.1007/s11858-010-0245-x

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437-475. https://doi.org/10.1023/A:1009084430926

Stoten, W. D. (2019). Metacognition, calibration, and self-regulated learning: an exploratory study of undergraduates in a business school. *Learning: Research and Practice, 5*(1), 24-47. https://doi.org/10.1080/23735082.2018.1491616

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving . *Journal of Educational Psychology, 82*(2), 306-314.

Tall, D. (2013). *How humans learn to think mathematically.* Cambridge University Press.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research.* SAGE Publications, Inc.

Thiede, K. W., Anderson, M. C., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66-73. https://doi.org/10.1037/0022-0663.95.1.66

Thomas, G. P. (2002). The social mediation of metacognition. In D. McInerny, & S. Van Etten (Eds.), *Sociocultural influences on motivation and learning* (Vol. 2, pp. 225-247). Information Age Publishing.

Thomas, G. P. (2003). Conceptualisation, development, and validation of an instrument for investigating the metacognitive orientation of science classroom learning environment: The Metacognitive Orientation Learning Environment Scale-Science (MOLES-S) . *Learning Environment Research*, 175-197. https://doi.org/10.1023/A:1024943103341

Thomas, G. P. (2006). An investigation of the metacognitive orientation of confucian heritage culture and non-confucian heritage culture science classroom learning environments in Hong Kong. *Research in Science Education, 36*, 85-109. https://doi.org/10.1007.s11165-005-3915-x

Thomas, G. P. (2013). The interview as a metacognitive experience for students: Implications for practice in research and teaching. *ASEJ, 43*(1), 4-11.

Thomas, G. P., & Anderson, D. (2014). Changing the metacognitive orientation of a classroom environment to enhance students' metacognition regarding chemistry learning. *Learning Environment Research, 17*, 139-155. https://doi.org/10.1007/s10984-013-9153-7

Thomas, G. P., & McRobbie, C. J. (2001). Using a metaphor for learning to improve students' metacognition in the chemistry classroom. *Journal of Research in Science Teaching, 38*(2), 222-259.

Thomas, G. P., & McRobbie, C. J. (2012). Eliciting metacognitive experiences and reflection in a year 11 chemistry classroom: An activity theory perspective. *Journal of Science Education and Technology, 22*, 300-313. https://doi.org/10.1007/s10956-012-9394-8

Thomas, G. P., Anderson, D., & Nashon, S. (2007). Development of an instrument designed to investigate elements of science students' metacognition, self-efficacy, and learning processes: The SEMLI-S. *International Journal of Science Education*, 1-24. https://doi.org/10.1080/09500690701482493

Thomas, G., & Anderson, D. (2012). Parents' metacognitive knowledge: Influences on parent-child interactions in a science museum setting. *Research in Science Education, 43*, 1245-1265. https://doi.org/10.1007/s11165-012-9308-z

Tobias, S., & Everson, H. (1996). *Assessing metacognitive knowledge monitoring.* College Board.

Tobias, S., & Everson, H. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw, *Issues in the measurement of metacognition* (pp. 147-222). Buros Institute of Mental Measurements and Erlbaum Associates.

Tobias, S., & Everson, H. (2002). *Knowing what you know and what you don't: Future research on metacognitive knowledge monitoring.* College Board.

Valdez, A. (2013). Student metacognitive monitoring: Predicting test achievement from judgment accuracy. *International Journal of Higher Education, 2*(2), 141-146. https://doi.org/10.5430/ijhe.v2n2p141

van der Stel, M., & Veenman, M. V. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences, 20*, 220-224. https://doi.org/10.1016/j.lindif.2009.11.005

van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merrienboer, J. J. (2013). Activation of inaccurate prior knowledge affects primary school students' metacognitive judgments and calibration. *Learning & Instruction, 24*, 15-25. https://doi.org/10.1016/j.learninstruc.2012.08.005

van Luit, J. H., & Kroesbergen, E. H. (2006). Teaching metacognitive skills to students with mathematical disabilities. In A. Desoete, & M. Veenman (Eds.), *Metacognition in mathematics education* (pp. 177-190). Nova Science.

van Overschelde, J. P. (2008). Metacognition: Knowing about knowing. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 47-71). Psychology Press.

van Velzen, J. (2016). *Metacognitive learning: Advancing learning by developing general knowledge of the learning process.* Springer International Publishing.

Veenman, M. V., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences . *Learning and Individual Differences, 15*, 159-176. https://doi.org/10.1016/j.lindif.2004.12.001

Veenman, M. V., Kok, R., & Blote, A. W. (2005). The relation between intellectual and metacognitive skills at the onset of metacognitive skill development. *Instructional Science, 33*, 193-211.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition Learning*(1), 3-14. https://doi.org/10.1007/s11409-006-6893-0

Veenman, M. V., Wilhelm, P., & Beishuizen, J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*, 89-109. https://doi.org/10.1016/j.learninstruc.2003.10.004

Verschaffel, L. (1999). Realistic mathematical modelling and problem solving in the upper elementary school: Analysis and improvement. In J. H. Hamers, J. E. Van Luit, & B.

Csapo (Eds.), *Teaching and learning thinking skills: Contexts of learning* (pp. 215-240). Swets & Zeitlinger.

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition, 28*, 277-296. https://doi.org/10.1016/0010-0277(88)90016-9

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition, knowing about knowing* (pp. 157-184). MIT.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*(4), 408-428. https://doi.org/10.3200/GENP.132.4.408-428

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*(4), 173-188. https://doi.org/10.1207/s15326985ep3004_2

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*, 327-353. https://doi.org/10.1016/s1041-6080(96)90022-9

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (2 ed., pp. 153-189). Erlbaum.

Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environment. *International Journal of Educational Research, 41*(6), 466-488. https://doi.org/10.1016/j.ijer.2005.08.012

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, & J. Dunlosky (Eds.), *Metacognition in educational theory and practice.* Erlbaum.

Winne, P. H., & Jamieson-Noel, D. L. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology, 27*, 551-572. https://doi.org/10.1016/s0361-476X(02)00006-1

Woodworth, R. S., & Thorndike, E. (1900). Judgments of magnitude by comparison with a mental standard. *Psychological Review, 7*, 344-355. https://doi.org/10.1037/h0064288

Yates, F. (1990). *Judgment and decision making.* Prentice-Hall.

Yau, M., Rosolen, L., & Archer, B. (2018). *2016-17 Student census: School report for grades 9-12.* Toronto District School Board.

Zabrucky, K. M. (2010). Knowing what we know and do not know: Educational and real world implications. *Procedia Social and Behavioral Sciences , 2*, 1266-1269. https://doi.org/10.1016/j.sbspro.2010.03.185

Zhao, N., Valcke, M., Desoete, A., & Verhaeghe, J. P. (2011). A multilevel analysis on predicting mathematics performance in Chinese primary schools: Implications for practice. *Asia- Pacific Education Researcher, 20*(3), 503-520.

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: judgment processes and accuracy constraints. *Educational Psychology Review, 20*(2), 191-206. https://doi.org/10.1007/s10648-008-9073-8

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses. *Contemporary Educational Psychology, 11*, 307-313. https://doi.org10.1016/0361-476(86)90027-5

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329-339. https://doi.org/10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*, 82-91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 1-37). Erlbaum.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166-183. https://doi.org/10.3102/0002831207312909

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299-315). Routledge.

Zimmerman, B. J., & Schunk, D. H. (2003). Albert Bandura: The scholar and his contributions to educational psychology. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Educational psychology: A century of contributions* (pp. 431-457). Erlbaum.

Appendix A

Conceptualisation of Calibration Over the Past Four Decades

	Source	Proposed Definition of Calibration
1.	Fischhoff, Slavic, & Lichtenstein, 1977	Calibration is a term used to describe the correlation between the degree of confidence the learners have about their performance and their actual performance.
2.	Lichtenstein, Fischhoff, & Phillips, 1982	Calibration reflects the degree to which individuals' judgments about their understanding, capability, competence, or preparedness correspond to the understanding, capability, competence, or preparedness they actually manifest.
3.	Glenberg, Sanocki, Epstein, & Morris, 1987; Wagenaar, 1988	Calibration is a measure of the relationship between the confidence in performance and the accuracy of performance.
4.	Yates, 1990	Calibration is a measure of the degree to which individuals' judged ratings of performance correspond to their actual performance.
5.	Horgan, 1990	Calibration is defined as the accuracy with which students can predict their own performance.
6.	Keren, 1991	Calibration is the degree of fit between individuals' judgment of performance and their actual performance.
7.	Bjorkman, 1992; Cooke, 1991; Sharp, Cutler & Penrod, 1988	Calibration is a measure of the relationship between individuals' level of confidence in their knowledge and their actual performance.
8.	Glenberg & Epstein, 1985; Schraw & De Backer Roedel, 1994; Stone, 2000; Weingardt, Leonesio, & Loftus, 1994	Calibration refers to the congruence, alignment, match between students' metacognitive judgments and their performance.
9.	Lin & Zabrucky, 1998	Calibration is defined as the degree to which judgments of performance accurately reflect actual performance.

10. Gravalia & Gredler, 2002	Calibration is the extent of congruence between students' estimates of their capabilities and actual performance.
11. Jamieson-Noel & Winne, 2003	Calibration is the degree (accuracy) and direction (bias) of discrepancy between students' perception about learning and actual properties of learning.
12. Winne, 2004	Calibration refers to the degree to which learners' judgments about some feature of a learning task deviates from an objectively or externally determined measure of that feature.
13. Maki, Shields, Wheeler, & Zacchilli, 2005	Calibration or absolute accuracy assesses the precision of a confidence judgment compared to the performance on a criterion task.
14. Nietfeld et al., 2005; Zimmerman & Moylan, 2009	Calibration is the degree to which individuals' self-efficacy about performance matches their performance.
15. Alvarez & Emory, 2006	Calibration is an executive function that allows people to monitor their own performance and judge what they do and do not know.
16. Nietfeld, Cao, & Osborne, 2006a	Calibration is the degree to which individuals' perception of performance corresponds with their actual performance.
17. Nietfeld, Enders, & Schraw, 2006b	Calibration is a term used in the study of metacognition monitoring that describes the relation between task performance and a judgment about that performance.
18. Hacker, Bol, & Keener, 2008b	The degree of correspondence between one's perception of performance and actual performance is referred to as calibration.
19. Pieschl, 2009	Calibration refers to the degree of accuracy of learners' perceptions of their own performance.

20. Boekaerts & Rozendaal, 2010	Calibration describes the relationship between task performance and a judgment about that performance.
21. Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010	Calibration refers to how accurately individuals can predict how well they will do on a task prior to starting it and how well they judge their performance after completion.
22. Bol & Hacker, 2012	The difference between perceived and actual performance is referred to as calibration.
23. Valdez, 2013	Calibration refers to students' ability to accurately match their confidence of success with the actual outcome of a test.
24. Ots, 2013	The extent of agreement between a prediction about performance regarding a certain task and the actual performance
25. Rinne & Mazzocco, 2014	Calibration is the alignment of accuracy and confidence.
26. Crane, Zusho, Ding, & Cancelli, 2017	Calibration refers to learners' ability to monitor their own comprehension and performance.
27. Gutierrez, 2017	Calibration refers to the relation between criterion task performance and a judgment about that performance.
28. Gutierrez, 2017	Calibration is an outcome from the Feelings of Knowing (FOKs) judgments.
29. Gutierrez et al., 2017	Calibration is the process by which learners convey what they know and do not know about a topic.
30. Rutherford, 2017a	Calibration is the accuracy of one's judgments about one's performance
31. Stoten, 2019	Calibration is the relationship between students' perception of their performance and their actual performance.

Appendix B

Research Gaps as Directions for Future Research

Source	Research Gaps as Direction for Future Research
Lundeberg & Fox, 1991; McCormick, 2003; Winne, 2004	The majority of calibration studies have been conducted in laboratory settings. Although they have strong internal validity, reasons exist to believe that the generalisation of some of the findings to naturalistic contexts such as classrooms may not be possible.
Barnett & Hixon, 1997	The need exists for more mixed methods studies to direct research towards more successful interventions.
Hacker, Bol, Horgan, &Rakow, 2000	The need exists for intervention to calibrate students.
Nietfeld & Schraw, 2002	Strategy training for monitoring accuracy has not been done before.
McCormick, 2003	Given the importance, researchers now face the challenge of increasing monitoring accuracy in actual classroom practice. It is not clear if the results of lab settings can be generalised to classroom learning.
Zimmerman & Schunk, 2003	Intentional and systematic interventions are necessary to improve students' monitoring accuracy.
Nietfeld, Cao, & Osborne, 2005	Previous studies have been done in lab settings, as one-time testing sessions, which makes generalising problematic due to the lack of ecological validity. In laboratory-style studies, the period between learning and testing is considerably shorter than in real classroom learning. There is much more to calibration than what has been investigated so far. About 32% of judgement accuracy has been explained by students' explanatory styles, which means that calibration can be influenced by a more global self-concept of ability than what the actual performance data suggests. This may be one reason why calibration is often resistant to change.
Isaacson & Fujita, 2006	Future research should examine the following: Does students' monitoring change over time? What factors affect monitoring skills? Can these skills be taught to low-achieving students?

Nietfeld, Cao, & Osborne, 2006a	Little evidence exists on how improvement in calibration relates to complex representations of classroom material over an extended period of time. Little work has been done examining the change in calibration in relation to important affective factors such as self-efficacy, except for the work of Hacker, Bol, Horgan, & Rakow (2000), but they have also pointed out the lack of classroom studies over an extended period of time.
Chen & Zimmerman, 2007	Little is known about how self-efficacy beliefs affect post-performance judgements.
Hacker, Bol, & Keener, 2008b	Calibration researchers should conduct research in naturalistic settings of classrooms across grade levels and subject areas. The augmentation of quantitative data with rich qualitative data in mixed methods designs is also recommended. Nearly all the classroom studies we studied are quantitative. Identifying factors that contribute to the accuracy of calibration judgements remains a fertile area for research. The current gaps include the following: a) Nearly exclusive use of undergraduate students in calibration research has left us with limited understanding of other ages. b) The need exists for longitudinal studies to uncover the developmental aspects of calibration. c) The need exists for understanding the role of students' beliefs and social factors within the classroom setting. d) Domain specificity must be considered. e) The stability of calibration can possibly be attributed to the stability of personality traits or stable social factors. f) Researchers must replicate studies and consider social factors. g) Explanatory style and motivation may explain why calibration is resistant to change.
Dignath & Buttner, 2008	Research gap exists as to whether interventions are more effective in high school than elementary school.
Hacker, Bol, & Bahbahani, 2008a	Moving from the lab to classroom is necessary to understand cognition in the context of natural purposeful activities. Very few classroom studies exist. The impact of students' beliefs on calibration has not

	been tested. We recommend formulating interventions that develop accurate calibration and then testing their effectiveness in 'real-life' classroom contexts.
Pieschl, 2009	The exact nature of metacognitive judgements is still unclear. Some assume them as snapshots and others as stable.
Parkinson, Dinsmore, & Alexander, 2010	No studies exist on comparing the reliability and utility of item-by-item versus overall confidence judgements. Calibration has rarely been investigated in a domain-specific context, and whether calibration accuracy can be improved is a question that has not been definitively answered. 80% of the studies on calibration are not situated in academic domains such as mathematics or science, and 67.5% use convenience samples of undergraduates; hence, the developmental aspects of calibration are not reflected in research.
Parkinson, Dinsmore, & Alexander, 2010	In examining calibration, factor of self has been ignored.
Zabrucky, 2010	The reasons why the calibration of low-achieving students does not improve are not known. Previous interventions have not been helpful for them.
Miller & Geraci, 2011	Policies for numeracy and literacy exist but not learning to learn. The need exists to ensure that the needs of students who enter secondary school without necessary learning-to-learn skills are met. Developing self-regulated learning skills makes the high school experience more effective, less stressful, and ultimately more rewarding.
Salter, 2012	No research has studied the degree to which calibration improves or remains stable over time or across consecutive judgements of confidence.
Hadwin & Webster, 2013	Research gap exists as to what kind of interventions can be implemented to improve students' judgement accuracy, how can calibration be validly and reliably measured, and does one mode of rating or analysis have superiority.
Alexander, 2013	The need exists to clarify whether the positive relation between monitoring and performance still holds in a

	natural setting. It is not known if the findings in one subject can be generalised to others.
Roderer & Roebers, 2013	In most studies, only one test is used, so it is a snapshot of students' monitoring skills. Most studies in calibration are conducted in controlled lab-style settings, where tests are designed to elicit maximum monitoring. Tests are also administered by people unknown to students with minimal resemblance to a real classroom. The importance of natural setting and ecological validity is ignored.
Valdez, 2013	Current views have suggested that it is time to investigate calibration in more authentic settings such as classrooms, using authentic assessment and performance measures.
Stoten, 2019	Until recently, an overwhelming proportion of researchers have employed a quantitative approach to conduct research on calibration.

Appendix C

Inconclusive or Contradictory Results in Calibration Research

	Study	Mixed Findings
	Conceptualisation:	
1.	Alexander, Dinsmore, Parkinson, & Winters, 2010	Contradictory results may be due to the lack of a clear conceptual definition and grounding within a theoretical framework.
2.	Hacker, Bol, & Keener, 2008b	Contradictory results are found when different measures of calibration are used (e.g., predictions versus postdictions).
3.	Nelson & Narens, 1990	Metacognition is a stable psychological construct.
4.	Bol, Hacker, O'Shea, Allen, 2005; Hacker, Bol, Horgan, & Rakow, 2000; Thomas & McRobbie, 2012; Stoten, 2019	Metacognition is a dynamic socio-cultural construct.
	Malleability of Calibration:	
5.	Bol & Hacker, 2001; Bol, Hacker, O'Shea, Allen, 2005; Koriat, 1997; Nietfeld, Cao, & Osbourne, 2005; Nietfeld & Schraw 2002	Improving calibration accuracy is difficult or not durable.
6.	Glenberg, Sanocki, Epstein & Morris, 1987; Gutierrez de Blume, 2017; Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osbourne, 2006a; Schraw, Potenza, & Nebelsick-Gullet, 1993; Yates, 1990	Different types of intervention can lead to significant improvement in calibration accuracy.
7.	Bol & Hacker, 2001; Bol, Hacker, O'Shea, Allen., 2005; Nietfeld, Cao, & Osbourne, 2005	Three classroom studies showed that calibration tends to remain stable despite feedback and practice.
8.	Gutierrez & Schraw, 2015	Studies show that calibration accuracy is malleable and can be improved.

Reflection:	
9. Nietfeld, Cao, & Osbourne, 2006a	Reflection and instruction on self-assessment are found to be effective for all students.
10. Hacker, Bol, Horgan, & Rakow, 2000	Reflection and instruction on self-assessment are effective only for high achievers.
Feedback:	
11. Callender, Franco-Watkins, & Roberts, 2016; Stone, 2000; Zimmerman, 1990	Feedback is effective in improving calibration accuracy.
12. Schraw, Potenza, & Nebelsick-Gullet, 1993	Feedback is not effective in improving calibration accuracy.
13. Stone, 2000; Zimmerman, 1990	In lab, generally, feedback improves calibration
14. Nietfeld, Cao, & Osbourne, 2005	In a classroom setting, feedback does not improve calibration.
Calibration in a Classroom:	
15. Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osbourne, 2006a	Modest gain in calibration accuracy has been achieved in classroom studies.
16. Bol & Hacker, 2001; Bol, Hacker, O'Shea, Allen, 2005; Nietfeld, Cao, & Osbourne, 2005	Negligible gain in calibration accuracy has been achieved in classroom studies.
17. Bol & Hacker, 2001	Less gain was achieved in calibration accuracy when the teacher reviewed with students than when students reviewed without a teacher.
Confidence Judgments:	
18. Pajares & Kranzler, 1995	Confidence judgements are measures of self-efficacy.
19. Hadwin & Webster, 2013	Confidence judgements are indicators of monitoring. They are not a measure of self-efficacy because they are not about students' general ability to learn.

Appendix D

Invitation Letter to Parents/Guardians and Informed Consent Form for Students under 18 Years of Age

Dear Mr. / Mrs. ………… Parent/Guardian of………… Date: ______________

My name is Behnaz Herbst and this semester, I will be ……….'s mathematics teacher in grade 11. In addition to teaching, I am also a PhD Candidate at the University of Alberta. The focus of my research is on metacognitive calibration; my intention is to close the gap between students' self-estimations of their performance and their actual performance in mathematics, and I am writing to request ……………'s participation in research. I have enclosed a Research Information Sheet to provide you with some background information about my research, but I am also writing to invite you and ………. to a Research Information Night, which I will be hosting on Monday, Feb. 8 in our school library from 7 PM to 9 PM. In this meeting, I will explain my research and answer any questions you have.

All students in the course are invited and may take part in research related activities that I will introduce as a component of my regular pedagogy in teaching the course. However, only the data of students with a signed parental consent form will be included in research. I have asked Ms. Edwards in the Main Office to be in charge of collecting the signed consent forms, because I do not wish to know who the research participants are. In the enclosed Research Information Sheet, you will find a description of the research procedures in more detail, but to explain briefly, students will begin by completing a questionnaire. During the following 10 weeks, they will estimate their grades on quizzes and tests, will graph their estimated grade against the actual grade, and will plan how to close the gap between their estimated and actual grades. The final stage of the study includes another questionnaire and follow-up interviews. The last day of the study is Friday, May 13, 2016, however, I will conduct the interviews during the first two weeks of July.

Participation is on a voluntary basis. Students in my course have the option of joining another section of the course with the same content and assessment but no research component, and those who decide to remain in the course I am teaching, can always terminate their participation and withdraw their research-related data at any stage of the study even after May 13. I do not wish to be informed if my students participate or withdraw, so they will only have to inform Ms. Edwards in the Main Office about the withdrawal from the study. Students also have the option of staying in my class without participating in research. The plan for this study was reviewed for its adherence to ethical guidelines by the University of Alberta Research Ethics Board (REB 2) and the District School Board External Research Review Committee (ERRC). To participate in research, the District School Board requires a signed parental consent form. I will bring the consent forms to the Research Information Night on Monday, Feb. 8, but for your convenience, I will attach the form to this letter, as well.

Should you require more information, please feel free to E-mail me at behnaz.herbst@... or call the Mathematics Department at ……… Should you need a translator, please let me know and I will make the necessary arrangements.

I would like to welcome ………. to the course and I look forward to personally meeting and greeting you both on Feb. 8.

Thank you for considering my request and with warm regards,

Behnaz Herbst

Parent/ Guardian Informed Consent Form (Parents of Students under 18 Years of Age)

I, _______________________________ have read the accompanying Research Information Sheet and give my informed consent for my son/daughter to participate in the research study titled *A Classroom Study of the Development of Mathematics Students' Metacognitive Calibration*, conducted by Ms. Behnaz Herbst, my son/daughter's mathematics teacher, which will take place on school premises between February 1, 2016 and May 13, 2016, with the exception of the follow-up interviews, which will be conducted after the semester ends.

I consent to my child (Please check each item you consent to) participation in the following research-related activities,

- o Complete two questionnaires, one at the beginning and the one in the end of the study
- o Participate in calibration sessions and activities
- o Participate in performance estimation and graphing
- o Participate in completing student notecards
- o Participate in brief audio-recorded interviews, after the semester ends.

In agreeing for my child to take part in this study, I understand that:

- Students are under no obligation to participate
- The classroom teacher will not know who the participants are until after the semester ends
- Even after giving my consent for my child to take part, he/she may discontinue his/her participation without penalty at any time and his research material will be withdrawn at the end of the semester.
- Information that is provided will be treated as confidential. Direct quotes from my child may be used in research-related presentations and publications only after his/her name and other personal identifiers have been removed
- Without identifying individual students, data from this research will be used for the purpose of a PhD dissertation and research reports may be used for academic and professional presentations for the advancement of mathematics education.

I understand that I am under no obligation to consent to my child's participation in this study and that I and/or my child can withdraw consent from the study at any point in time after which any information or data linked to my child as an individual will be withdrawn from the study.

____

(Print Name)

______________________________________ __________________________

(Signature) (Date)

Please note that two copies of this form will be provided. One copy should be signed and returned, and the other copy should be kept for your records. Should you have any questions or concerns please feel free to contact the teacher-researcher directly at …………...

Appendix E

Invitation Letter to Students, Informed Assent Form for Students under 18 Years of Age, and Informed Consent Form for Students over 18 Years of Age

Dear ……………. Date: ________________

My name is Ms. Herbst and I will be your mathematics teacher in the grade 11 course that you are taking this semester (MCR3U). I am writing to invite you to participate in the research component of this course, which will focus on students' self-estimation of test performance and will attempt to close the gap between students' perceptions of performance and their actual performance on mathematical tasks, such as tests and exams, with potentially positive effects on students' strategic efforts and study behaviour. I am also writing to invite you and your parents to a Research Information Night which I will be hosting on <u>Monday, Feb. 8 in our school library from 7 PM to 9 PM</u>. In this meeting, I will explain my research and answer any questions you and your parents might have.

All students in the course are invited to participate in this research. The enclosed Research Information Sheet will provide you with a description of research procedures and what students will do, but to explain briefly, students will begin by completing a brief questionnaire. During the following 10 weeks, they will estimate their grades on quizzes and tests, will illustrate their expected grades against the actual grades on a graph, and will plan strategies to improve accuracy. In the final stage of the study, students will complete another questionnaire and participate in audio-recorded interviews. Students who are 18 years old can sign their own consent, otherwise, they need to sign an assent form and will need a consent form signed by their parents/guardians. For your convenience, I have attached both forms to this letter. Ms. Edwards in the Main Office is in charge of collecting the signed consent forms. I am asking all my students to go through the research activities, because I do not wish to know who the actual research participants are and whose data will be included in the research. With exception of the interviews, the last day of data collection for this study will be Friday, May 13, 2016, and the last date for the withdrawal of data will be July 30, 2016.

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at 780 492 2615. Participation is on a voluntary basis. Right now, you have the option of joining another section of the course with no research component, and if you decided to remain in the course I am teaching, you can terminate your participation in research at any later time by simply letting Ms. Edwards in the Main Office know. You also have the option of staying in my class without participating in research.

I look forward to working with you this semester and will be pleased to explain my research to you on Research Information Night next Monday, Feb. 8 and answer any questions you might have.

Thank you for considering my request, and with best regards,

Sincerely,

Behnaz Herbst

Assent Form for Students under 18 Years of Age

Date: ___________

Please sign below to indicate your willingness to take part in the study described above and be sure to return the signed forms to Ms. Edwards in the Main Office.

I, _______________________________ have read the accompanying Research Information Sheet and give my informed consent to participate in the research study titled ***A Classroom Study of Mathematics Students' Metacognitive Calibration***, conducted by Ms. Herbst, my mathematics teacher, which will take place on school premises between February 1, 2016 and May 13, 2016, with the exception of the follow-up interviews, which will be conducted shortly after the semester ends.

I hereby agree to (Please check each item you consent to):

- o Complete two questionnaires, one at the beginning and the one in the end of the study
- o Participate in calibration sessions and activities
- o Participate in performance estimation and graphing
- o Participate in completing student notecards
- o Participate in brief interviews, after the semester ends.

In agreeing to take part in this study, I understand that:

- I am under no obligation to participate
- The classroom teacher will not know who the participants are until after the semester ends
- Even after giving my consent to take part, I may discontinue my participation without penalty at any time and may also withdraw the research material that was already collected by contacting Ms. Edwards in the Main Office.

- Information that is provided will be treated as confidential. Direct quotes from me may be used in research-related presentations and publications without any reference to my name or any other personal identifiers.
- Without identifying students, research reports will be used for the purpose of a doctoral dissertation and for academic and professional presentations.

I understand that I am under no obligation to consent to participation in this study and that I can withdraw consent from the study at any point in time after which any information or data linked to me as an individual will be withdrawn from the study.

____

(Print Name)

_____________________________ _______________________

(Signature) (Date)

Dear student: Please note that two copies of this form will be provided. One copy should be signed and returned to Ms. Edwards in the Main Office (NOT to the classroom teacher), and the other copy should be kept for your records. Should you have any questions or concerns about this study, please feel free to contact the teacher-researcher directly at ……………....

Consent Form for Students over 18 Years of Age

Date: __________________

Please sign the form below to indicate your willingness to take part in the study described above and be sure to return the signed forms to Ms. Edwards in the Main Office.

I, _______________________________ have read the accompanying Research Information Sheet and give my informed consent for to participate in the research study titled *A Classroom Study of Mathematics Students' Metacognitive Calibration*, conducted by Ms. Herbst, my mathematics teacher, which will take place on school premises between February 1, 2016 and May 13, 2016, with the exception of the follow-up interviews, which will be conducted shortly after the semester ends.

I hereby agree to (Please check each item you consent to):

- o Complete two questionnaires, at the beginning and in the end of the study
- o Participate in calibration sessions and activities
- o Participate in performance estimation and graphing
- o Participate in completing student notecards
- o Participate in brief audio-recorded interviews, after the semester ends.

In agreeing to take part in this study, I understand that:

- I am under no obligation to participate

- The classroom teacher will not know who the participants are until after the semester ends
- Even after giving my consent to take part, I may discontinue my participation without penalty at any time and may also withdraw the research material that was already collected by contacting Ms. Edwards in the Main Office.
- Information that is provided will be treated as confidential. Direct quotes from me may be used in research-related presentations and publications without any reference to my name or any other personal identifiers.
- Without identifying students, research reports will be used for the purpose of a doctoral dissertation and for academic and professional presentations.

I understand that I am under no obligation to consent to participation in this study and that I can withdraw consent from the study at any point in time after which any information or data linked to me as an individual will be withdrawn from the study.

__________________________________ __________________________

(Print Name) (Date)

____

(Signature)

Dear student: Please note that two copies of this form will be provided. One copy should be signed and returned to Ms. Edwards in the Main Office (NOT to the classroom teacher), and the other copy should be kept for your records. Should you have any questions or concerns about this study, please feel free to contact the teacher-researcher directly at …..……..

Appendix F

Research Information Sheet and Contact Information

Title of the Research Study

A Classroom Study of the Development of Mathematics Students' Metacognitive Calibration

Timeline:

February 1, 2016 – May 13, 2016 (Interviews will be conducted shortly after the semester ends)

Background and Purpose

To be successful in academic subjects such as mathematics, students need to be able to self-evaluate their test performance. The accuracy of these evaluations, also called *metacognitive calibration* is considered important for learning, motivation, and achievement. In other words, students might decide not to spend much time and effort studying and practicing for a test, unless they can distinguish between what they can and cannot do mathematically. The present study is a fourteen-week-long classroom intervention, designed to teach grade 11 mathematics students a set of skills which will help them become more accurate in their self-evaluation of mathematical tasks.

Procedures

Phase I: During the first two weeks of the semester (Feb. 1-12, 2016) students will complete a questionnaire and receive some background information about calibration.

Phase II: Starting the third week of school, every Tuesday, students will receive a brief lesson on calibration. For example, these lessons may include "what is calibration and why is it important for students to be well calibrated?" Also, after each formal assessment such as a quiz or a test, and before it is graded, students will make an estimate of their grades. When the test is returned, they will reflect on the reasons why there might be a gap between the estimated and received grades, will graph the results, and will make plans for more accurate performance evaluation.

Phase III: During the last two weeks of the study (May 2-13, 2016) students will complete another brief questionnaire. Soon after the semester ends, the participants will take part in brief audiotaped interviews (cameras/ video-recording devices will NOT be used).

All the signed consent and assent forms should be directed to Ms. Edwards in the Main Office. The classroom teacher does not wish to know who the research participants (with a signed consent/assent form) are until the semester ends.

Benefits for Students

This intervention has several benefits for participating students:

- Participants will learn and practice a set of metacognitive calibration skills that are essential to successful task performance in mathematics.
- General metacognitive skills can be used in all subject areas. They are considered life-long learning skills that students can use throughout their lives.
- As students become more accurate in self-evaluation of task performance, their test performance in mathematics is expected to improve.
- By realizing their autonomy over their study behaviour and performance, students may recognize they are in charge of their grades and can perform better by expending more time and effort.

Privacy and Confidentiality

Personal records and research material relating to this study will be kept confidential. However, the research data will be used without identifying students. Participants' names will not be disclosed to any outside party and any report published about this study will not identify the participants. After data collection is completed, all names will be removed and replaced by Research ID Numbers and hence no part of this study or the research material may be traced back to the individual students. Research related data will be stored in secure locked filing cabinets in the Main Office at school, until the end of the semester.

Voluntary Participation

Although all students in the course are encouraged to participate in this study, participation is on a voluntary basis. Students who present a signed parent/guardian consent form to Ms. Edwards in the Main Office will be assigned a Research ID Number and while all the students in the course will go through the stages of the study together, only the research material of students with Research ID Numbers will be included in the data analysis.

Students have the option of joining a different section of the course with the same content and assessment but no research component and those who decide to remain in the course can withdraw their participation and their research-related data at any stage of the study until July 30. Students also have the option of staying in the course without participating in research. Other than Ms. Edwards who is in charge of the consent forms, nobody (not even the classroom teacher) knows who the participants are and students who wish to withdraw from the study are asked to do so without informing the classroom teacher. They are only required to inform Ms. Edwards in the Main Office.

Contact Names and Phone Numbers

With any questions or concerns, please contact:

Classroom Teacher-Researcher

Behnaz Herbst
Mathematics Department
(School Name)

(School Phone Number)
behnaz.herbst@.....

Research Supervisors

Dr. Florence Glanfield
Vice-Provost (Indigenous Programming & Research) &
Professor, Department of Secondary Education
University of Alberta

(780) 492 6492
glanfiel@ualberta.ca

Dr. Greg Thomas
Professor of Science Education
Department of Secondary Education
University of Alberta

(780) 492 5671
gthomas1@ualberta.ca

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board (REB 2) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at 780 492 2615.

Appendix G

Student Questionnaire #1

Name____________________ Date______________

Instructions:

i) Please read the following questions carefully and answer in detail, to the best you can, in your own words. There are no right or wrong answers.
ii) There are 4 questions in total. Use the back of the sheets should you need more space for your answers.
iii) Please ask me if you are not sure what the question is asking. Thank you!

1. Are you sometimes surprised by your score when you get a test or quiz back? Explain.

2. If there is sometimes a gap between the score you expected and the one you received, in your view, what are the reasons for this gap?

3. After writing a test, do you think you could accurately estimate your test score? If so, how do you think you will make this estimation and what would you base this estimation on?

4. Do you usually receive a grade which is more than what you expected, less than what you expected, or the same as what you expected? Why do you think this is the case?

Appendix H

Student Questionnaire #2

Name____________________ Date______________

Instructions: i) Please read the following questions carefully and answer in detail to the best you can, in your own words. There are no right or wrong answers.
ii) There are 4 questions. Use the back of this sheet should you need more space for your answers.
iii) Please ask me if you are not sure what the question is asking. Thank you!

1) You have been participating in educational research for the past 3 ½ months. Congratulations! What does 'calibration' mean to you? How do you make sense of it and what do you think it means or does?

2) In your experience, how did your calibration change and develop over time, if it did?

3) Each time that you noticed a gap between your estimated and actual scores, what did you think the main reasons were?

4) Each time that you made a performance estimation, how and based on what did you make your estimations?

Dear Students,

Please see the next page for three important questions. Thank you!

Three Important Questions

1) As an important component of this research, would you like to participate in a brief audio-recorded interview with me? No cameras will be used, and I will only audio record your answers. These interviews will take place in the first two weeks of July. Thank you!

   Please check the appropriate box:

   Yes, I am interested in participating in audiotaped interviews ☐

   No, thank you! ☐

2) After the completion of data analysis, would you be interested in receiving some feedback and recommendations regarding your calibration? Please check the appropriate box:

   Yes, I would like to receive feedback ☐

   No, thank you! ☐

3) After June 30, would you like the originals of your research material (questionnaires, notecards, and your calibration graph) to be returned to you? Please check the appropriate box:

   Yes, I would like the originals back ☐

   No, thank you! ☐

Appendix I

Ten Weekly Calibration Sessions, Modelling, and Reinforcement

Session # 1 Script Tue. Feb. 16, 2016

Title: Self-Evaluation of Task Performance Matters!

How important is it for a pilot to realize that he is flying too close to the summit? How crucial is it for a surgeon to know that she has almost had no training for an upcoming surgery? How about you?! How important is it for you; mathematics students to realize that you need to review some topics and solve some more problems to perform well on your upcoming test?

You are correct in thinking that these are very different scenarios, but in fact, they all rest on the same premise: the importance of self-evaluation of task performance; in all these scenarios, people need to be able to accurately evaluate their own performance (usually against some perceived standard), with some consequences to follow if they are unable to do so.

Back to the example of mathematics students, let's do some brainstorming and see why self-evaluation of performance might be important for students. (Discussion)

This Friday, you will write your first quiz. Remember that you will receive part-marks for relevant partial steps and expressions of mathematical thinking. After answering the questions, please make an estimation of your performance; that is your expected grade in this quiz. In the upcoming weeks, we will have more quizzes, tests, a midterm exam, and a final exam. After each one, you will be asked to make an estimate of your performance. For simplicity, we call all the quizzes and tests in this course 'Q*uests'* and will show them as Q1, Q2, Q3, and so on. Thanks!

Lesson #1: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

Working out a problem set on the board I make some mistakes, intentionally, but keep going. When the solution is complete, I give myself an estimation of 85% on the entire task and wait for comments and discussion. What are the underpinnings of accurate estimations in students' opinion? What are the consequences of inaccurate performance estimations?

Second, I ask some students to each choose a set of homework questions, solve them on the board, and give themselves a performance estimation percentage. We discuss *how* one should evaluate task performance in mathematics. This may help prepare the students for a similar activity on Friday and initiate some thought-provoking conversations.

Finally, students sit in 'groups of professionals'. In one group, the members are all doctors, others are groups of teachers, journalists, judges, nurses, lawyers, and so on. As they interact, one member pretends that she/he cannot accurately estimate her/his task performance on a certain task. The professional group members discuss how this issue may affect the whole group. This scenario intended to reinforce the importance of self-evaluation of task performance in social and professional settings.

Lesson # 2 Script Tue. Feb. 23, 2016

Title: What Does It Mean to Be Calibrated?

Last Tuesday, we talked about the importance of accurate performance estimation. To the extent that a person can estimate his or her own task performance accurately, we say that he or she is *calibrated*. Are you familiar with the word 'calibration'? What do you think it means? (Discussion)

Calibration is a term used in engineering and instrumentation. It is a comparison between a measurement which is the standard (accepted as correct) and one of unknown accuracy. Suppose you regularly weigh yourself at the gym, but it seems that the 'calibration' of the weighing scale is off because it reads ½ lb without any load. The manager promises she will get the maintenance crew to *calibrate the scale* and asks you to use the digital scale which is *calibrated*. What do you think being calibrated means for students to be well calibrated in terms of their task performance? (Discussion) What does it mean for students to be poorly calibrated? (Discussion)

Last Friday, you wrote a quiz, and on your quiz, you made a performance estimation. Here is your quiz! I have marked it. Is your estimated grade the same as your actual grade?

Please find the notecard I have attached to your quiz, and take a moment to answer two questions: Why do you think the grades differ and what would be a good strategy for you to use to close the gap?

You have also each received a personal calibration graph. It is a blank grid with your name on it. Please take a moment and point-plot your performance on your first quiz as Q1 on this graph. In order to do this, take your estimated grade in % as the x-value and take your actual grade in % as the y-value. Then point-plot the point Q1(x, y) on the calibration graph. Thank you!

Lesson #2: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- I put a question on the board and start solving it, while thinking aloud. I present a scenario in which I do not know how to solve this problem, but I am *aware* of the prospect of my poor performance on this task and so I spend some time asking students to help me solve the problem and looking at similar problems in the textbook. Then I ask: What do you think happened here? What was the person's weakness in this scenario? What was her strength in this scenario? (Discussion)

- Groups of two, one student acts as a well-calibrated mathematics student, pilot, lawyer, and so on, and the other one as a poorly calibrated member of the same group. They present a brief skit of their professional aspirations and plans to the class. What happens when people misjudge their own performance preparedness?

- Groups brainstorm and make suggestions as to how one can improve one's calibration accuracy and help others to achieve the same goal.

Lesson # 3 Script Tue. March 1, 2016

Title: Overestimation & Underestimation of Performance

Here is the story so far:

In our first lesson, we looked at the importance of accurate estimation of task performance, and in our second lesson we learned about the concept of *calibration*: To the extent that people can accurately estimate their own task performance, they are calibrated.

I am also returning your tests to you. Please take a moment to answer the two questions you have on your new notecards and point-plot your performance as Q2 on your personal calibration graph.

Have you wondered why I asked you to 'graph' your performance? Why do you think I asked you? (Discussion)

What do you make of that line that you all have on your calibration graphs; the line at 45° called the *line of perfect calibration*. What do you think it stands for? (Discussion)

What do you think it means to overestimate task performance? (Discussion)

What does it mean to underestimate task performance? (Discussion)

What is your status? Are you overestimating, underestimating, or neither? Are you perfectly calibrated?

Lesson #3: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- I invite two students to illustrate overestimation and underestimation of task performance, while they study for an important test together.

- I invite skits for brief enactment of overestimation and underestimation within the frame of schoolwork, but it can be in any subject area.

  In groups of three, each group chooses a certain profession. Within the group, one person demonstrates overestimation in preparing for a task, the second person demonstrates underestimation, and the third person demonstrates perfect calibration. They discuss the issues each faced and report back to the whole class.

Lesson # 4 Script   Tue. March 8, 2016

Title: What Are the Estimation Judgments Made of?

Just to refresh your memory here are the ideas we have talked about, so far:

- Self-estimation of task performance (and self-evaluation of task performance against an external standard) matters.
- Well-calibrated people estimate their performance with accuracy, while poorly calibrated people don't.
- There are two types of miscalibration: overestimation and underestimation. A person may overestimate himself/herself on one type of task and underestimate himself/ herself on another type.

When a person makes an estimation judgment, he or she makes that judgment based on 'something'. Please take a moment and discuss with each other or in groups how your estimations are made and based on what. What is the nature of these estimates? (Discussion)

Also, I am now returning your tests. Would be great if you could please take a moment to complete your notecards and point-plot this performance as the point Q3 on your personal calibration graphs!

The questions on students' reflection notecards asked two specific questions about the reasons for the potential gap between students estimated and achieved performance scores and the strategies students had to close the gap. By reminding the students of the previous calibration sessions, I intended to make the content and terminology more accessible to students and facilitate their writing of more detailed and elaborate explanations on their reflection notecards.

Lesson #4: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- I present a scenario in which I am basing my estimation of test performance in math on my previous performance on a history test and will invite comments.

- In groups of three using chart-paper and markers, each group will brainstorm and provide a list of the reasons given by group members for their estimations. For instance, are these estimations based on previous performance in math, previous calibration, previous performance in other subjects, personal effort, or time spent studying? What factors influence these estimation judgments? One of the students will do a synthesis of the data contributed by all groups.

- After group conversations, group members list on the board how they make their performance judgments. This may help students who are not sure how these judgments are made.

(March Break: The week of March14. School re-opens March 21)

Lesson # 5 Script Tue. March 22, 2016

Title: Role of Overestimation and Underestimation in Guiding Study Behaviour

Welcome back! Here is what we have talked about so far:

- Self-estimation of task performance (and self-evaluation of task performance against an external standard) matters.
- Well-calibrated people estimate their performance with accuracy, while poorly calibrated people don't.
- There are generally two types of miscalibration: overestimation and underestimation.
- An individual's estimation of task performance is made based on many factors.

Let me ask you this question: Andrew is a senior math student like yourselves and he is two days away from an important test. He has not studied any of the course topics before. The course content has been completely new and somewhat challenging to him. He has also missed many classes and extra help sessions and done neither any thinking about the topics nor any homework, but he thinks that he will get at least 85%. In a sense, he is ---------? Yes, he is overestimating his performance. My question for you is: Do you think that Andrew's estimation will affect his study behaviour for this test? Will you elaborate? (Discussion)

In this scenario Andrew 'does not know his stuff' so to say, but he thinks that he knows enough to perform at 85% level. Imagine a classmate of Andrew, Alida, a good math student who has been very conscientious. She has attended all the classes and done all the homework while engaging in mathematical thinking and reasoning and 'she knows her stuff' per se. Alida will make an estimation of about 65%. In a sense, Alida is ----------? Yes, she is underestimating her performance. How would this underestimation guide her study behaviour? (Discussion)

I am also returning your last test. Please spend a few minutes completing your notecards and point plotting Q4. Thanks!

Lesson #5: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- Class is divided into three groups and each group has a piece of large chart-paper and some markers. In this scenario, the members of the first group are all over-estimators, members of the second group are all under-estimators, and the members of the third group are all well calibrated. An important exam is approaching, and students have limited time. Each group will make their own list of the ways they will spend valuable study time.

- In groups, students brainstorm: How can a person *recognize* that he is overestimating or underestimating performance, and once this realization is achieved, how can he become better-calibrated?

- Groups brainstorm: In what ways might miscalibration influence study behaviour and come up with a list. A compiled list will be posted on the wall and on class website.

Lesson # 6 Script	Tue. March 29, 2016

Title: I Understand Calibration, but What Is Metacognition?

Hi everyone! Here is the story so far, just to refresh your memory:

- *Self-evaluation of task performance matters!*
- *If you evaluate your performance with accuracy, you are 'well calibrated'.*
- *'Miscalibration' means that you either overestimate or underestimate your performance.*
- *People base their performance judgments on different factors.*
- *Calibration guides students' strategic study behaviour and affects their academic success.*

In teaching you about calibration, I have sometimes used its correct (and long) name: *metacognitive calibration*. This is because calibration is a component of a much larger concept, called metacognition. But what is metacognition? Metacognition means to be aware of your thinking and also to regulate your thinking. For instance, if you have ever repeated to yourself or written down a phone number because you did not want to forget it you are metacognitive. The reason is that not only you are aware of your thinking processes (for instance, you know that you have to somehow manage to hold the phone number in your memory), but also you can think of control strategies, such as repeating or writing down the phone number to manage and control your thinking processes. Now, think about a mathematical function you want to graph, so you might begin by factoring the defining equation to find the zeros or do some point plotting to get a general shape. This is 'cognition' at work. But you may also think: Have I seen this function before? How difficult would it be to graph this function? Do I know enough to graph this function? Here, you are thinking *about your own cognition*. This is called metacognition. In other words, the subject of your thinking can be a graph, or a movie, or a cloud in the sky, and it can also be 'your thoughts'. In the simplest, metacognition has been defined as Thinking about your own thinking, as well as the monitoring and control of your thinking. Got it? Questions?

So, doing math engages your cognition, but thinking being aware of your thoughts while doing math is *metacognition* (e.g., maybe I should slow down while reading so I understand better – this is called a monitoring process). Calibration is all about monitoring because to be well calibrated means that you can distinguish between the 'known' and 'not-yet known' concepts and direct your time and effort toward the 'not-yet-known' concepts.

I am also returning your tests. Please complete your notecards and plot the point Q5 on your personal calibration graph. Thank you!

Lesson #6: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

Due to the importance of monitoring for successful mathematical problem-solving, in the following scenarios I combine the topic of metacognitive calibration with mathematical problem-solving scenarios for the benefit of my students.

- I model how problem-solving can go wrong without proper monitoring and possibly result in miscalibration. I choose a non-routine problem and solve it without monitoring my steps, explaining what can go wrong, and then once again with monitoring my steps. I remind the students that the accuracy of monitoring judgments is called calibration and that monitoring processes guide successful mathematical thinking and problem-solving steps. Different textbook problems that are familiar and relevant for students will be used.

- I review the different stages of a metacognitive problem-solving heuristic, which I taught students at the beginning of the semester: read more than once, think/analyse, plan your solution, implement your solution, and verify. I will provide examples and connect the overseeing of these steps to monitoring and calibration.

- I give students a set of non-routine math questions that require some metacognition to solve. Students work together, and we discuss what is different about these questions. One example of non-routine questions is: *Find as many Pythagorean triplet of integers as you can. An example is {3,4,5}, because* $3^2 + 4^2 = 5^2$. *Make a conjecture about a general rule, which can be used to obtain all these Pythagorean triplets of integers.*

Lesson # 7 Script Tue. April 5, 2016

Title: Metacognition in Mathematical Tasks

OK, who can tell me something about metacognition? What else? Why does it matter, in your view? Why is it important for mathematics students – and all students for that matter – to be 'aware' of their thinking and monitor and control their thinking while on academic tasks?

We talked about problem-solving and you may agree with me that metacognition (some call it reflection or mindfulness, but the scientific term is metacognition) is the essence of mathematical problem-solving and without metacognition, successful problem-solving might be highly improbable. This is mainly because senior mathematics students frequently face novel mathematical tasks that require attention and mental effort. What do you think this means in terms of monitoring? (Discussion)

Now that you are becoming more aware and more knowledgeable about monitoring and controlling of your thoughts, I also want to mention that the human brain does not multitask (Medina, 2008). Is this news to you? There is a simple reason for it: Your brain's neural circuits that are engaged in one task such graphing a parabola are not the ones that will be engaged in a completely different task such as understanding what is happening on a TV show that you are watching simultaneously. So, what your brain does is 'task-switching' (Medina, 2008), which means that you will be slower and more prone to making errors than if you would have focused on one task at a time. Metacognitive individuals have figured this out. They are very aware of their thoughts and can bring them back when they stray away. Since metacognition is so crucial to mathematical problem-solving, let's spend a few minutes to build on what we did last week, and I will show you how the stages of problem-solving work when we use metacognition. Let me take up your homework questions *metacognitively*!

I am also returning your tests. Please complete your notecards and point plot Q6 on your personal calibration graphs. Thank you!

Lesson #7: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- Groups of three or four, choose your partners, choose a problem, and take a large sheet of chart paper. Consult with each other and decide which problem-solving steps you are going to use to solve the problem. Some examples include reading, thinking, analyzing, and planning. Identify the steps that you and your group take and solve the problem.

- Groups of two students to act as Student and Teacher. The teacher reviews the problem-solving steps, gives the student a problem to solve, and will supervise the solution carefully, then they switch places. This way, we can take up some of the homework questions.

- Write one of the homework questions or any question that you like on a piece of paper. Then provide your own heuristic for problem-solving. In other words, explain in words, what your method will be (e.g., First I do …. because … and then I will ……). You may also solve the problem if you want to.

Lesson # 8 Script   Tue. April 12, 2016

Title: Importance of Calibration Graphs

Hi everyone! Just to refresh your memory, in the last two lessons we focused on

Lesson #6: Metacognition means thinking about your own thinking, as well as monitoring and controlling of your thinking,

Lesson #7: In problem-solving, metacognition helps us see if the method works, if we are moving in the intended direction and whether we are getting meaningful answers.

Do you have any questions regarding these two topics?

Today, we will focus on the calibration graphs you have been doing. These graphs, as you know, demonstrate how close you are to 'perfect calibration' but how? (Discussion) Can this continuous self-generated feedback help students adjust their time and study behaviour to the difficulty of the tasks they are facing at school? How? (Discussion)

As you know, calibration graphs can show you whether you are overestimating or underestimating your performance. How do you think this information might be helpful? (Discussion)

Let me return your test. Please spend a few minutes completing your notecards and point plotting the point Q7 on your calibration graphs. Thanks!

Lesson #8: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- In groups of 3 or 4, make a calibration graph for an imaginary student, whose Q1, Q2, and so on show overestimation at the beginning but she gradually becomes more accurate as time goes by. What would this students' calibration graph look like? Name the points Q1, Q2, and so on and be sure to include the Line of Perfect Calibration. Can you make a few recommendations to help student improve his/her calibration?

- Yesterday, you helped a student become better calibrated. Now go back to your groups and come up with a graph for another imaginary student, who in an under-estimator. What would his/her graph look like? Can you make some recommendations to help him/her?

- We have focused on calibration graphs. These graphs are an excellent source of visual feedback about performance. Today, individually or in your groups, comment on what you have learned or can learn *from working with these graphs*, and from what you or your friends have experienced in this regard. Any comments and suggestions are welcome.

Lesson # 9 Script Tue. April 19, 2016

Title: Thinking fast and Slow on Mathematical Tasks

Calibration is about our mental 'decision making' and without getting too technical and boring, I want to briefly tell you about how behavioural and cognitive psychologists think that our minds work. To borrow from Daniel Kahneman who won the 2002 Nobel Prize for behavioural economics, the human mind works by incorporating two systems that he calls System 1 and System 2. System 1 is fast and intuitive. It is the first thing that comes to your mind. It works without mental effort and can only answer easy questions. System 2 is the slow system. It is rational and logical and uses reason and mental effort. When you look at a math question and immediately, sometimes without even reading the question to the end, you say: *Oh, I know this,* and you start solving the problem, that is System 1 at work, because System 2 is slow and will not decide so quickly. It will go through the cycle of logical thinking and will make sure you understand the problem and have a logical way to solve the problem before giving you the green light. My point is: 1) the first answer that comes to your mind in doing mathematics is not usually the best one unless you are an expert on the topic and have a great deal of experience and 2) when you come up with a quick answer for a math question, always be suspicious that maybe your System 1 has tricked you by replacing a more difficult question with an easy one it knows how to answer. Do you want to try an example? Listen to my voice and think about this problem and see *what immediately comes to your mind*. Write it on a piece of paper just for yourself:

Together, a pen and a pencil cost $1.10. The pen costs one dollar more than the pencil. How much does the pencil cost? [Pause] Yes, the first response is $0.10, which is wrong. The correct answer is $0.05. Let's try another question:

All roses are flowers. Some flowers fade quickly. Therefore, some roses fade quickly. Right or wrong? It is wrong, although it sounds right to System 1. In estimating your performance, doing math, or any other thinking activity we need to put our System 2 to work. Please also complete your new notecards and plot point Q8 on your calibration graphs. Thanks!

Lesson #9: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- We talked about how the human mind works and my point was to bring to your conscious attention that logical thinking is not a given. It requires attention and mental effort. Therefore, as far as mathematics questions are concerned, unless you are an expert on the subject, the first answer that comes to your mind is most probably wrong. Decision making affects your calibration accuracy and your mathematical performance. Have you ever had a gut feeling which went wrong? I will tell you about one of mine then in groups of 4, please share one of yours!
- In groups, make a list of what you think System 1/System 2 do. I give you some examples.
- With a friend, tell each other what comes to mind first, then work the problems out:

a) *It takes 5 machines, 5 minutes, to make 5 rulers. How long would it take 100 machines to make 100 rulers? Would it be 100 minutes or 5 minutes?*

b) *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? Would it be 24 days or 47 days?*

Lesson # 10 Script Tue. April 26, 2016

Title: Where to next?

Thank you all for accompanying me on this journey of learning about calibration! Let's look at a summary of the previous nine lessons (I will review, briefly). You know so much now and I hope you will use this knowledge moving forward. Let me share with you that in the 1980s, educational researchers demonstrated that without any other factors being involved, students who set a goal/expectation for themselves showed significant and spontaneous academic progress. This is important and I encourage you to *always set a goal and work toward it in all subject areas, especially when you study for a test or an examination.* Also, remember System 1 and System 2 from now on. After studying/reviewing a portion of the content, which is going to be on a test, if you had a 'hunch' or a 'good feeling' that you knew enough and could stop studying, you should know that it is certainly your System 1 talking.

My hope is that you will always spend some time thinking logically (using System 2) about how much effort you have expended and what it takes to be successful on a task before deciding whether you are ready for your test or need to study some more. In decision making, you should be patient, take your time, and depend on logic and evidence, rather than basing your decisions on your feelings at the time. Now I need a person with a reasonably good handwriting to go to the board and I want each of us to contribute by telling him/her to write down one thing that they learned during this study.

Also, please complete your notecards and point plot Q9, which I returned, on your calibration graphs. Thank you!

Lesson #10: Modelling and Reinforcement

Ideas for Classroom Scenarios and Activities

- Remember that you cannot improve something that you are unaware of. Study habits are fundamentally important to your academic success. These habits strongly influence your decisions as to how to expend time, attention, and other cognitive resources on academic tasks toward successful task performance. Let's work in groups of 3 or 4. Everyone will contribute by writing down what their main study habits are and how they can improve it. Here, we will only focus on mathematics.

- An important goal of this study was to bring to your attention that you are in control of your thinking, study habits, and mathematics performance scores. What do you think this means and what do you think about your own autonomy and authority? (Discussion) What do you think affects your grades in math? (Discussion)

- You may find it interesting to know that although calibration research in psychology goes back to the year 1900 (more than a century ago), more than 80% of the contemporary research on calibration is not situated in an academic domain such as mathematics or science, but rather resides in professional and judicial areas. In groups of 4 or 5 make lists of what in your opinion calibration research can do for other fields and professions such as medical sciences, law enforcement, air and marine navigation, business, economics, management, engineering, and so on. This speaks to the importance of calibration for task performance in any professional or academic area that you might be interested in. Use Google if you want.

- General metacognitive skills are lifelong learning skills. How do you think these skills (e.g., planning, monitoring, evaluation) might be applied to your other academic courses to help you prepare for tests and exams?

- It is sometimes said that the best way to learn a topic is to teach it. Do you think you can teach someone about metacognition? Try and let me know.

Appendix J

The Calendar Representing the Intervention-Related Activities

FEBRUARY 2016

SUBJECT___Mathematics_MCR3U___ ________

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SAT/SUN
WEEK 1			3 **New Semester** Welcome & Introduction	4 Building context for the study, answering questions	5 Building context for the study, answering questions	6/7
notes						
WEEK 2	8 Questionnaire #1	9 Calibration Lesson + Monitoring strategy training	10 Questionnaire #1 (If missed) + Monitoring strategy training	11 Questionnaire #1 If missed) Monitoring strategy training	12 Assessment #1 Students choose questions + estimate performance scores	13/14
notes						
WEEK 3	15 Return test & Take up+ Calibration graphs + Reflection notecards	16 Calibration lesson + Monitoring strategy training	17 Monitoring strategy training	18 Monitoring strategy training	19 Assessment #2 Students choose questions + estimate performance scores	20/21
notes						
WEEK 4	22 Return test & Take up + Calibration graphs + Reflection notecards	23 Calibration lesson + Monitoring strategy training	24 Monitoring strategy training	25 Monitoring strategy training	26 Assessment #3 Students choose questions + estimate performance scores	27/28
notes						
WEEK 5	29 Return test & Take up + Calibration graphs + Reflection notecards					
notes						

MARCH 2016

SUBJECT Mathematics MCR3U

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SAT/SUN
WEEK 5		1 Calibration lesson + Monitoring strategy training	2 Monitoring strategy training	3 Monitoring strategy training	4 Assessment #4 Students choose questions + estimate performance scores	5/6
notes						
WEEK 6	7 Return test & Take up + Calibration graphs + Reflection notecards	8 Calibration lesson + Monitoring strategy training	9 Monitoring strategy training	10 Monitoring strategy training	11 Assessment #5 Students choose questions + estimate performance scores	12/13
notes						
WEEK	14	15 MARCH	16	17 BREAK	18	19/20
notes						
WEEK 7	21 Return test & Take up + Calibration graphs + Reflection notecards	22 Calibration lesson + Monitoring strategy training	23 Monitoring strategy training	24 Monitoring strategy training	25 Review & reinforcement of calibration lessons	26/27
notes						
WEEK 8	28 Return test & Take up + Calibration graphs + Reflection notecards	29 Calibration lesson + Monitoring strategy training	30 Monitoring strategy training	31 Monitoring strategy training		
notes						

APRIL 2016

SUBJECT Mathematics MCR3U

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SAT/SUN
WEEK 8					1 Assessment #6 Students choose questions + estimate performance scores	2/3
notes						
WEEK 9	4 Return test & Take up + Calibration graphs + Reflection notecards	5 Calibration lesson + Monitoring strategy training	6 Monitoring strategy training	7 Monitoring strategy training	8 Assessment #7 Students choose questions + estimate performance scores	9/10
notes						
WEEK 10	11 Return test & Take up + Calibration graphs + Reflection notecards	12 Calibration lesson + Monitoring strategy training	13 Monitoring strategy training	14 Monitoring strategy training	15 Assessment #8 Students choose questions + estimate performance scores	16/17
notes						
WEEK 11	18 Return test & Take up + Calibration graphs + Reflection notecards	19 Calibration lesson + Monitoring strategy training	20 Monitoring strategy training	21 Monitoring strategy training	22 Assessment #9 Students choose questions + estimate performance scores	23/24
notes						
WEEK 12	25 Return test & Take up + Calibration graphs + Reflection notecards	26 Monitoring strategy training	27 Monitoring strategy training	28 Monitoring strategy training	29 Assessment #10 Students choose questions + estimate performance scores	30///31
notes						

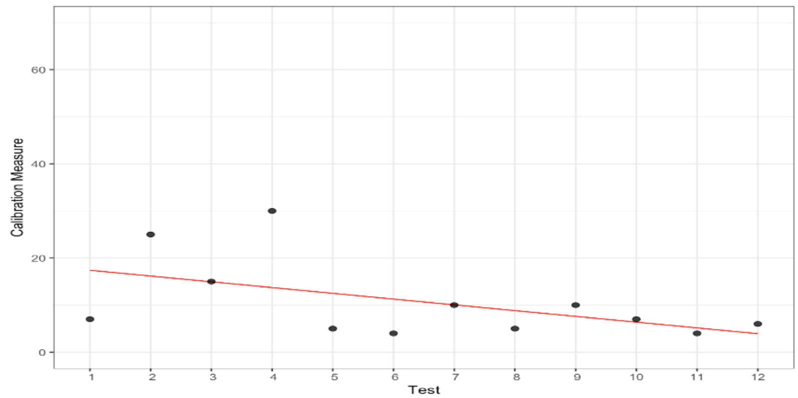
MAY 2016

SUBJECT Mathematics MCR3U

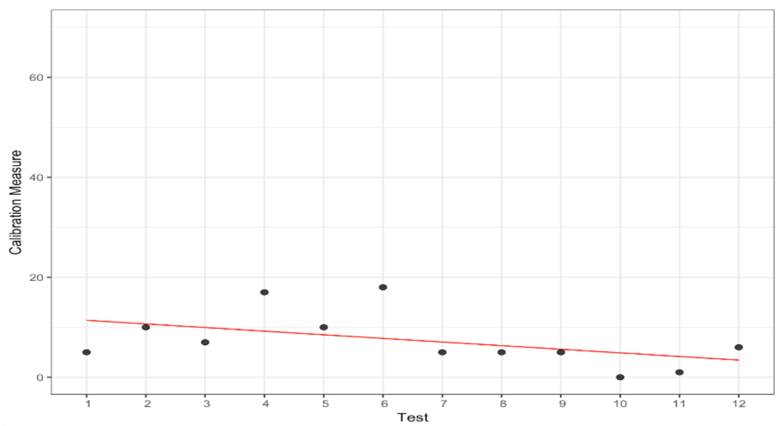
	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SAT/SUN
WEEK 13	2 Questionnaire #2	3	4 Questionnaire #2 (If missed)	5	6 Assessment #11 Students choose questions + estimate performance scores	7/8
notes						
WEEK 14	9 Return test & Take up + Calibration graphs + Reflection notecards	10 Monitoring strategy training	11 Monitoring strategy training	12 Monitoring strategy training	13 Saying thanks & closing remarks Assessment #12 is the final exam	14/15
WEEK notes						

Appendix K

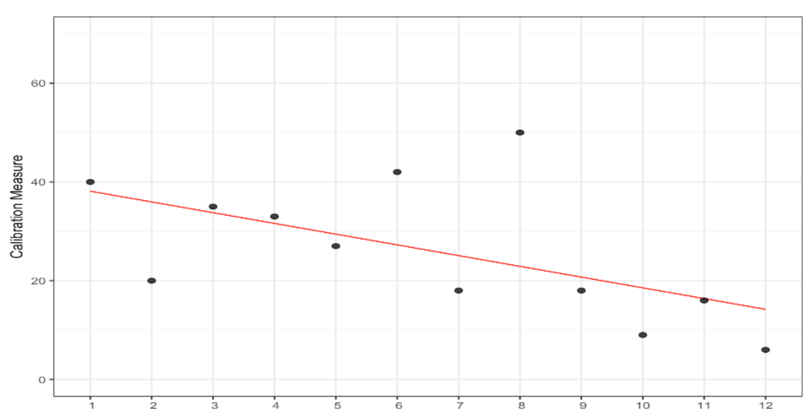
Students' Individual Calibration Graphs



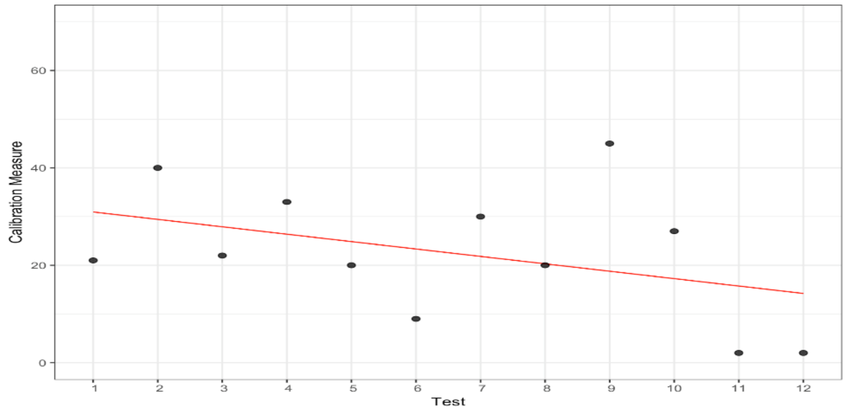
1. Abrar's calibration graph.



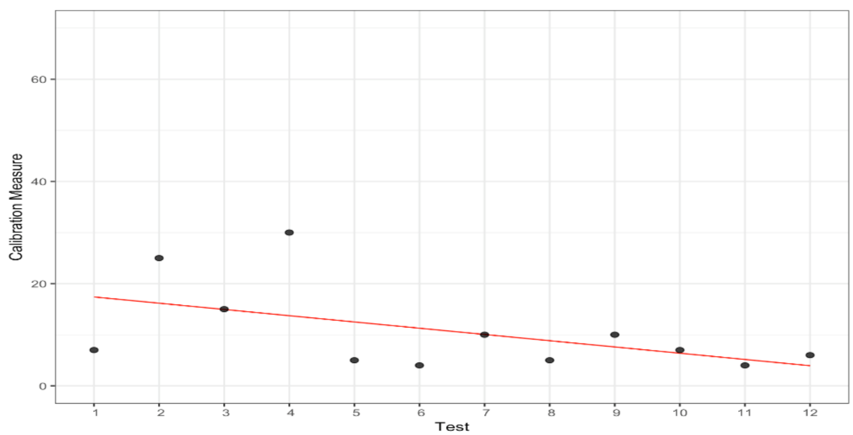
2. Adnan's calibration graph.



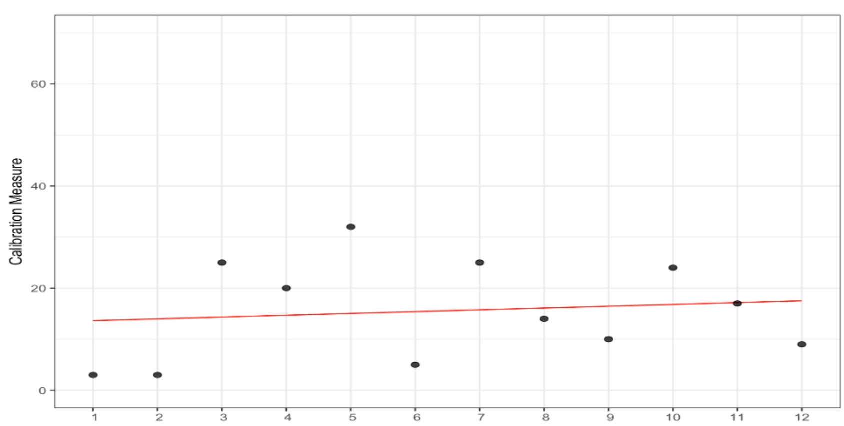
3. Andreas' calibration graph.



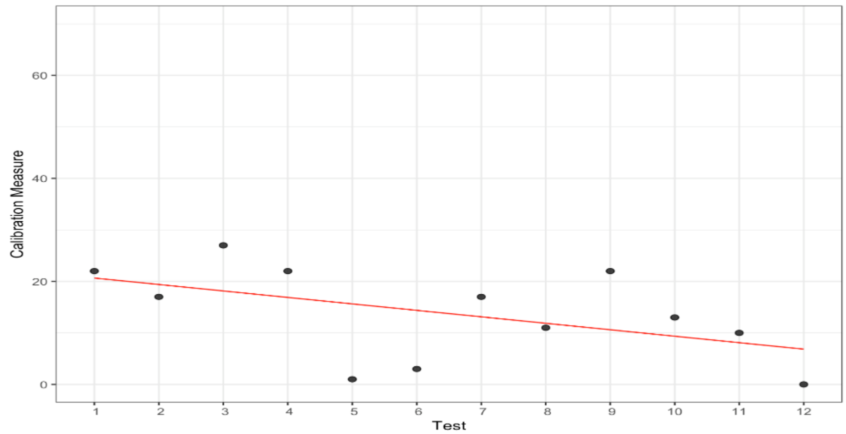
4. Anjani's calibration graph.



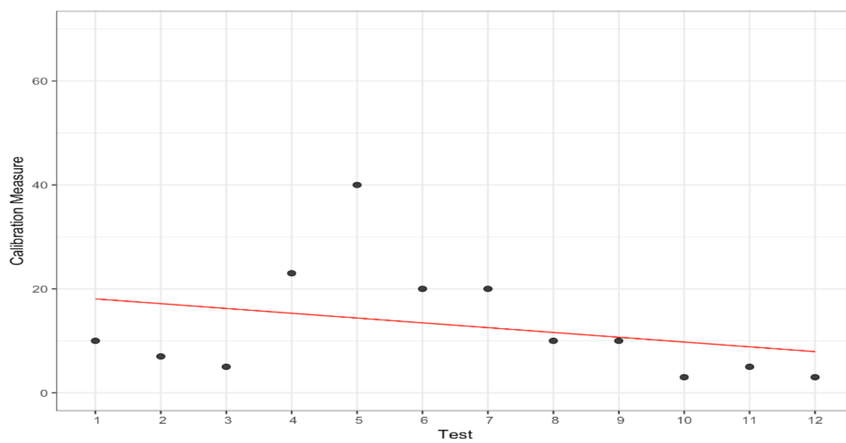
5. Anna's calibration graph.



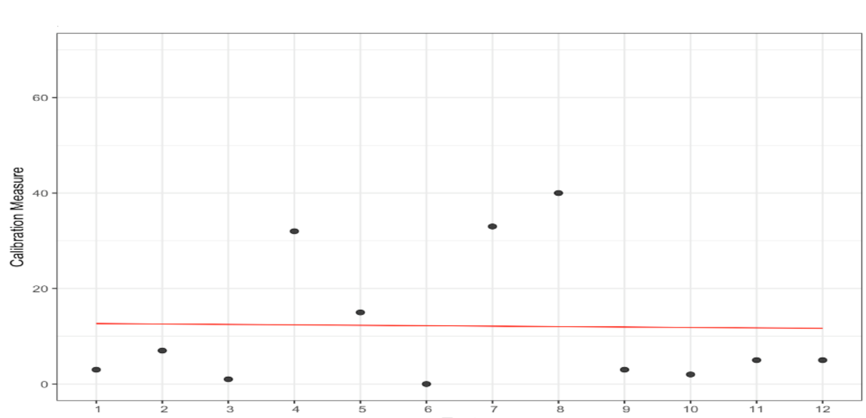
6. Bilal's calibration graph.



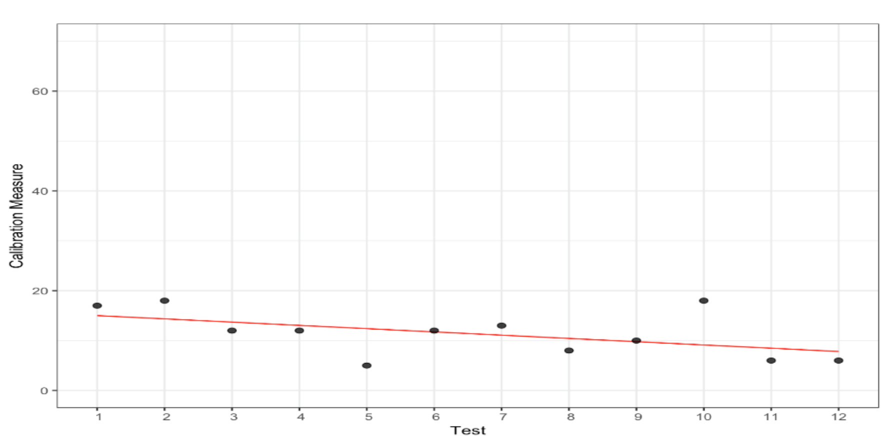
7. Fang's calibration graph.



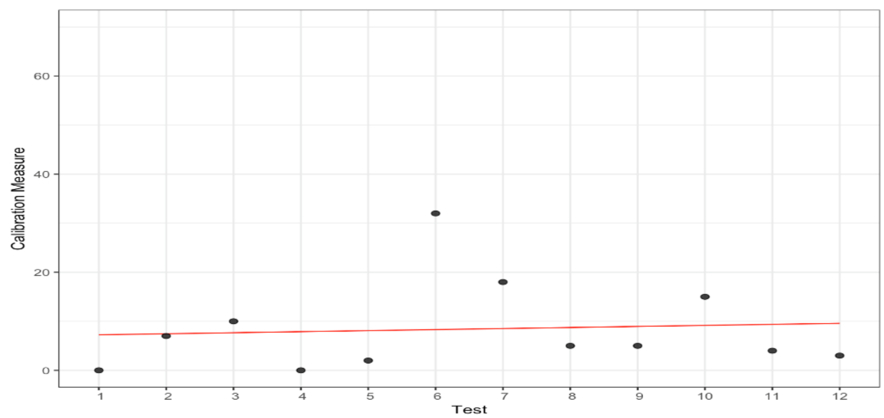
8. Faraz's calibration graph.



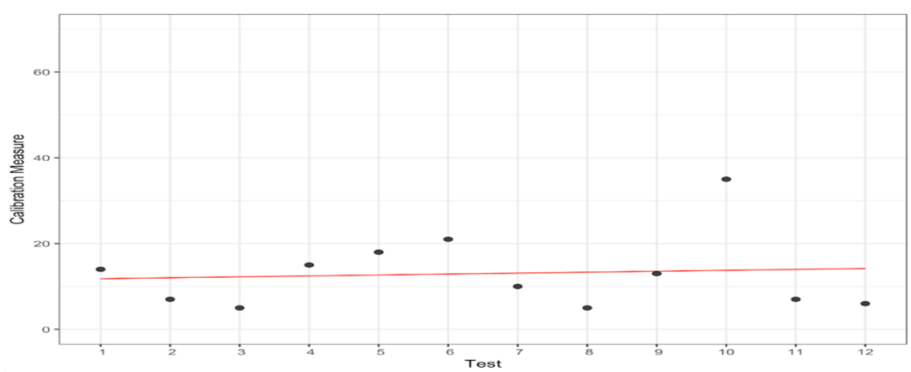
9. Kristian's calibration graph.



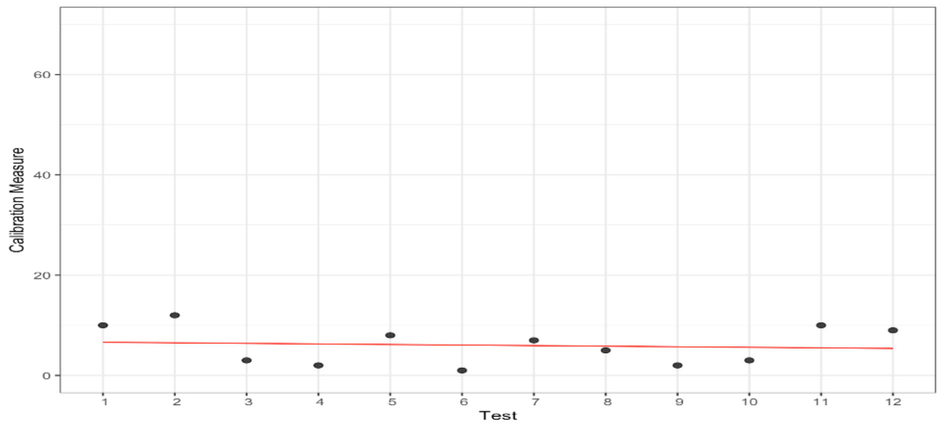
10. Leah's calibration graph.



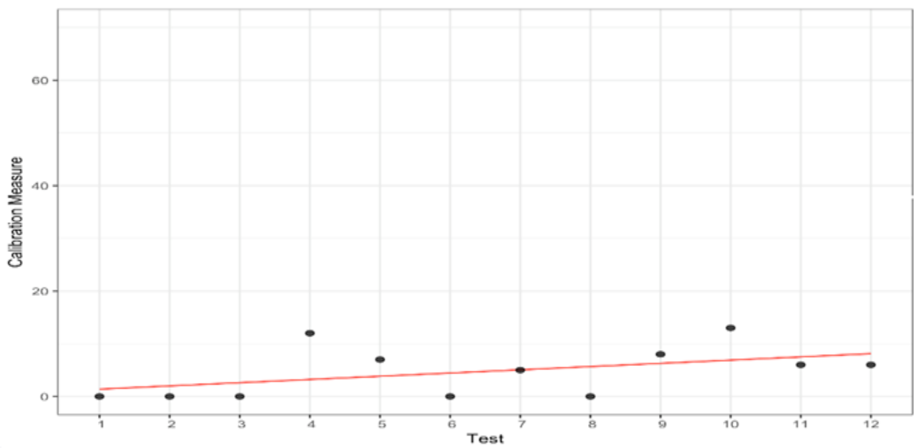
11. Li Min's calibration graph.



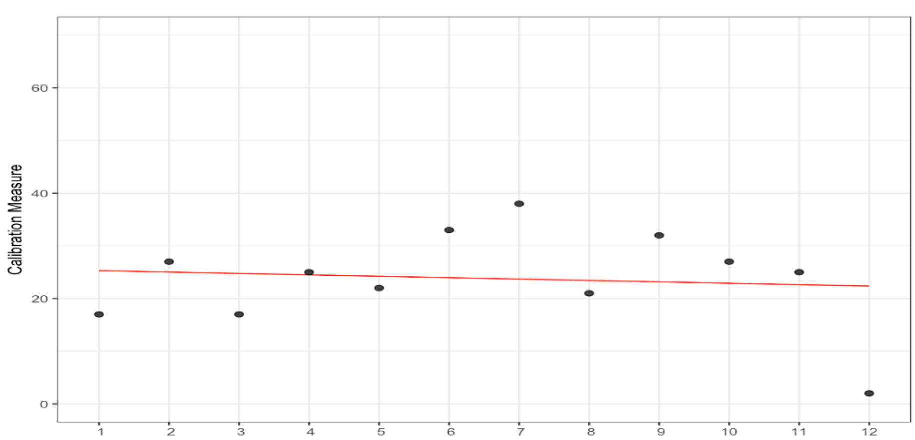
12. Liu Wei's calibration graph.



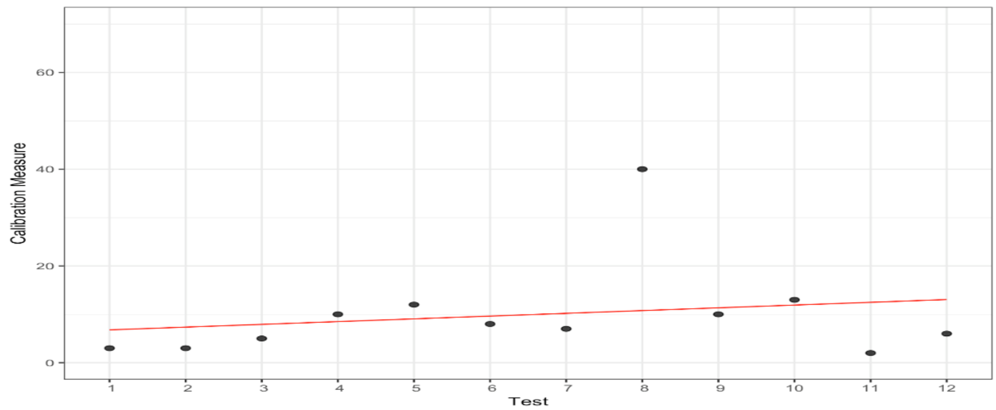
13. Lusine’s calibration graph.



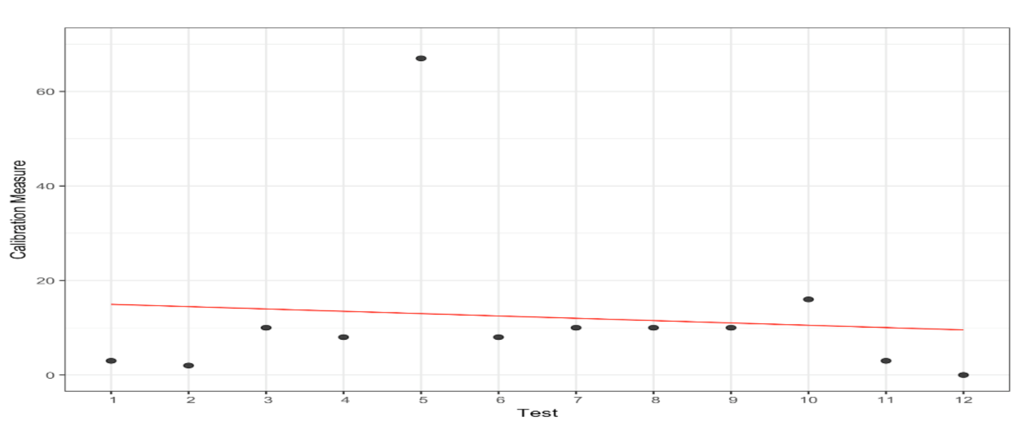
14. Marel’s calibration graph.



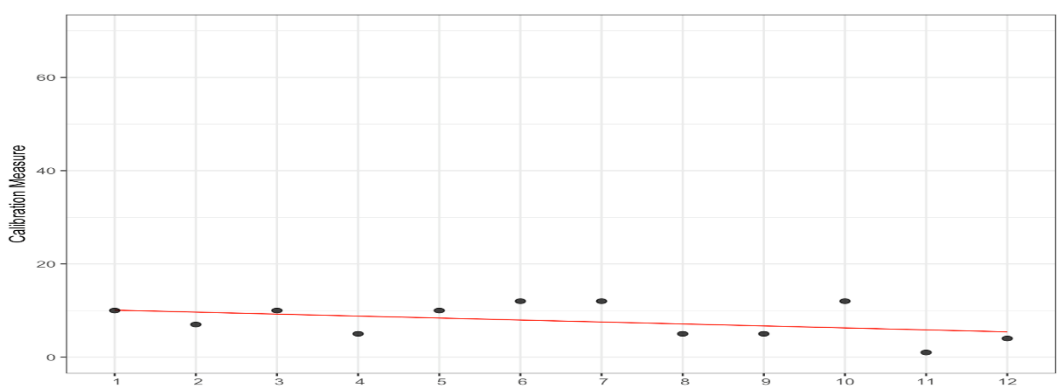
15. Neha’s calibration graph.



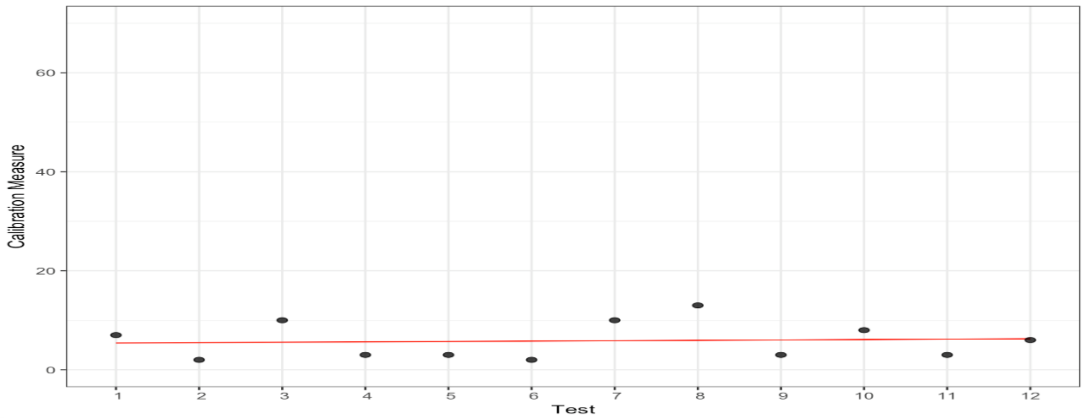
16. Rabia’s calibration graph.



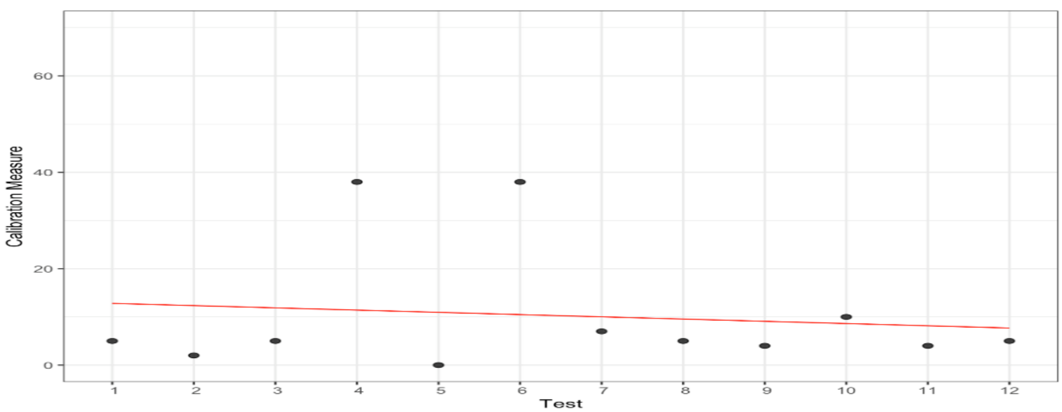
17. Sheng’s calibration graph.



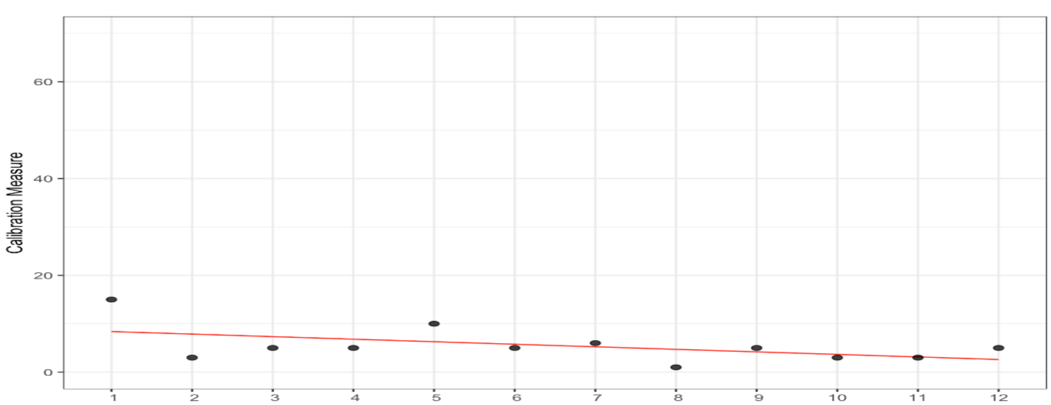
18. Shirin’s calibration graph.



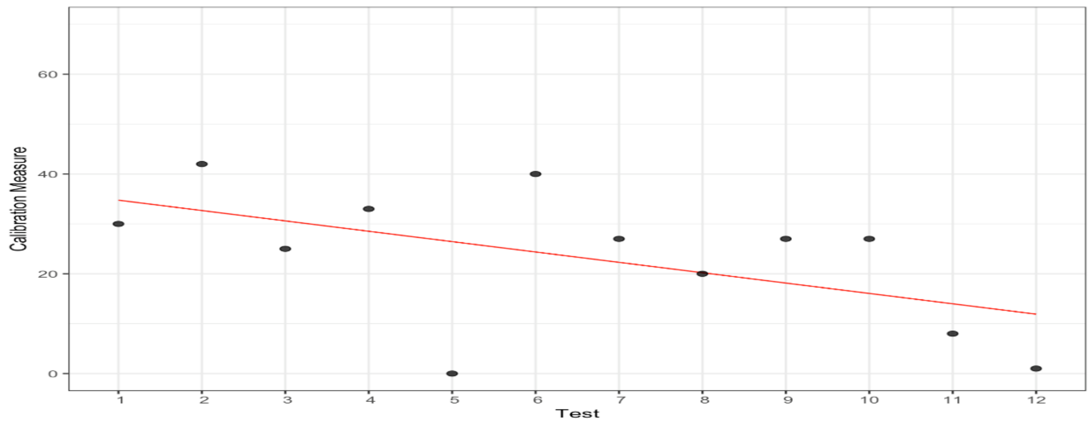
19. Sun Xin’s calibration graph.



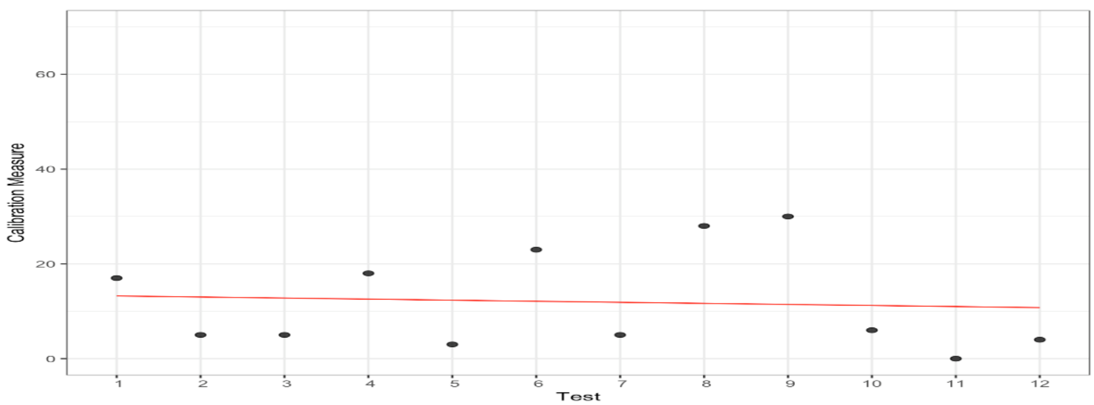
20. Tahir’s calibration graph.



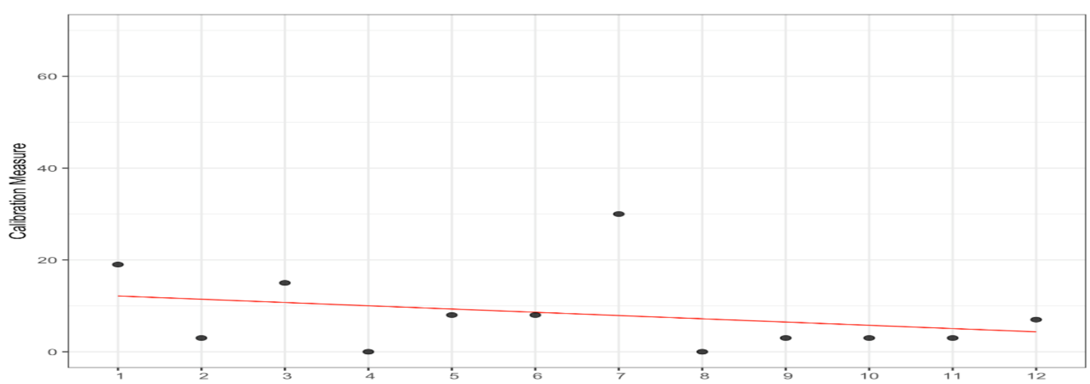
21. Wang’s calibration graph.



22. Yasir’s calibration graph.



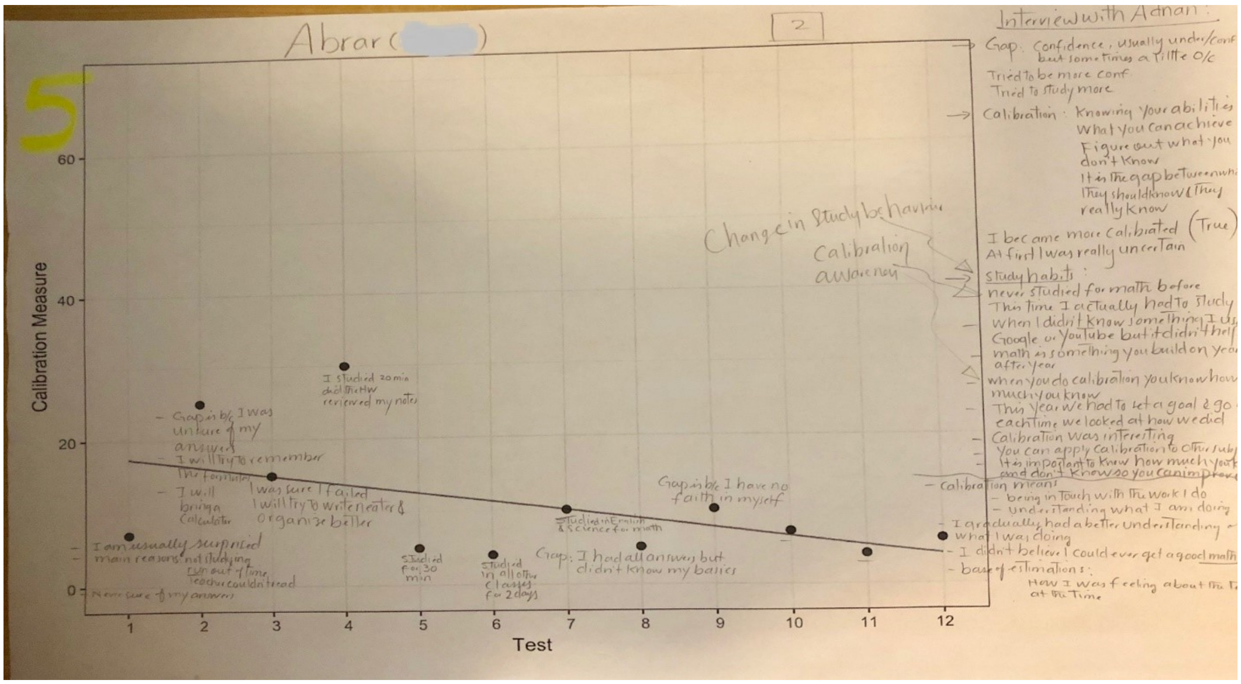
23. Zhang’s calibration graph.



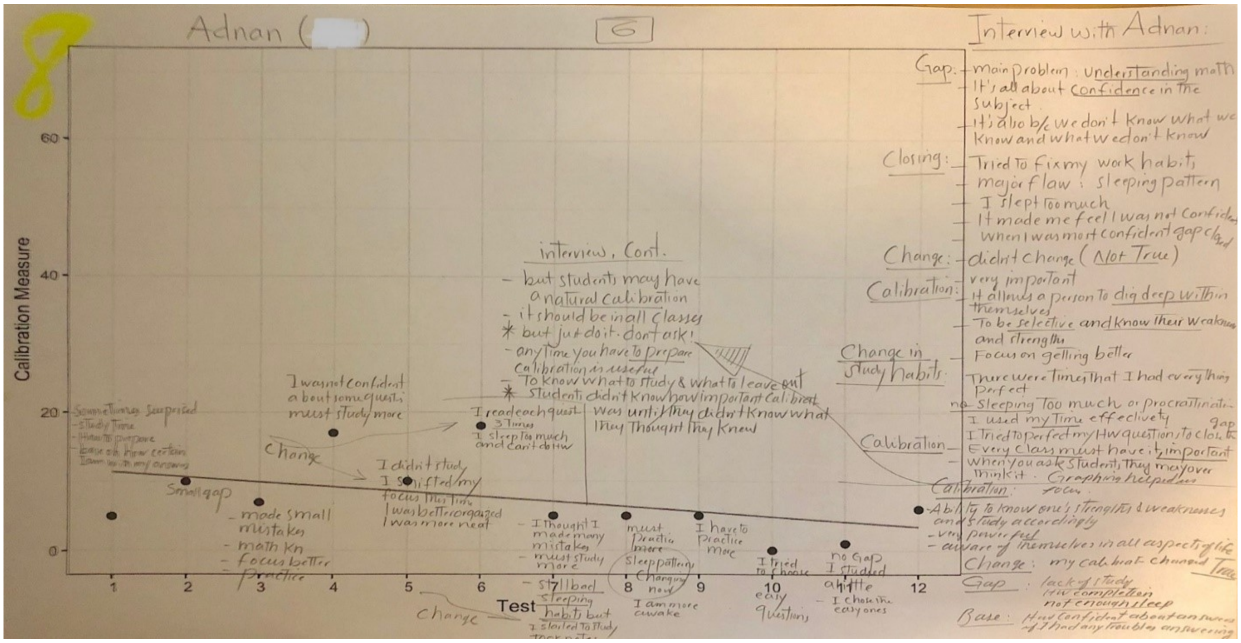
24. Zishan’s calibration graph.

Appendix L

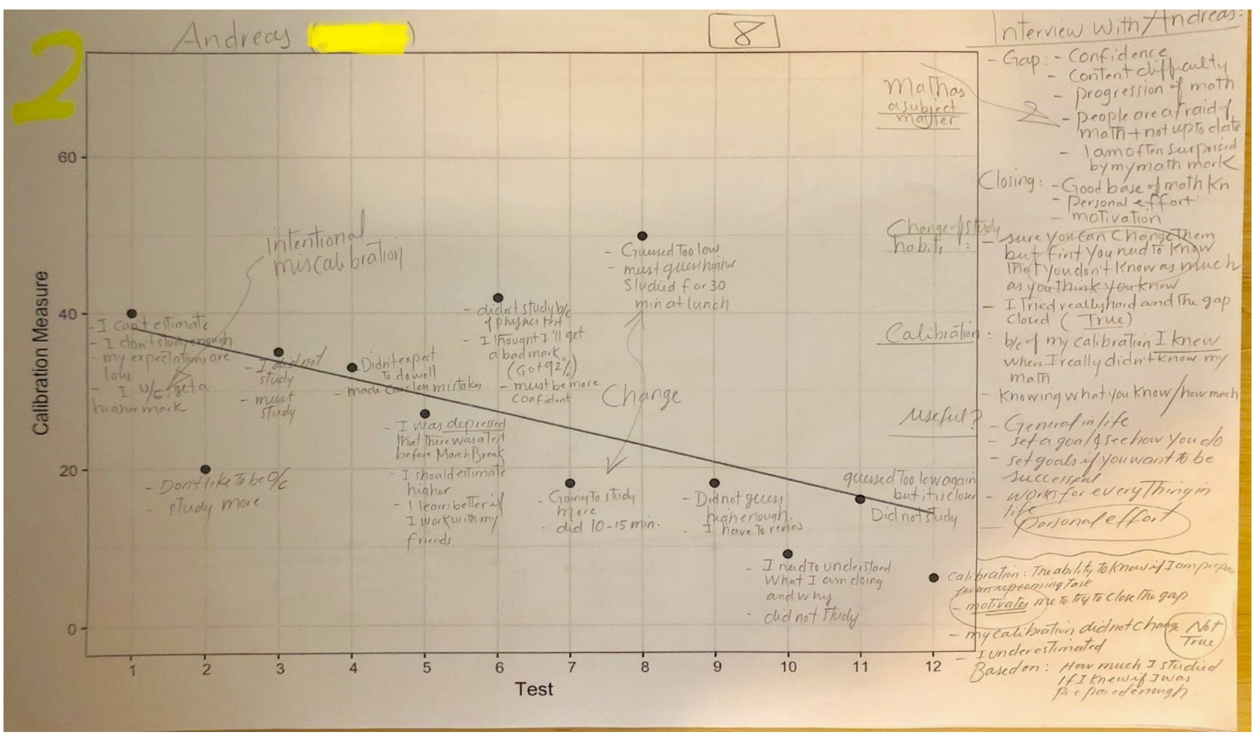
Seven Samples of Individual Student's Blended Data Analysis



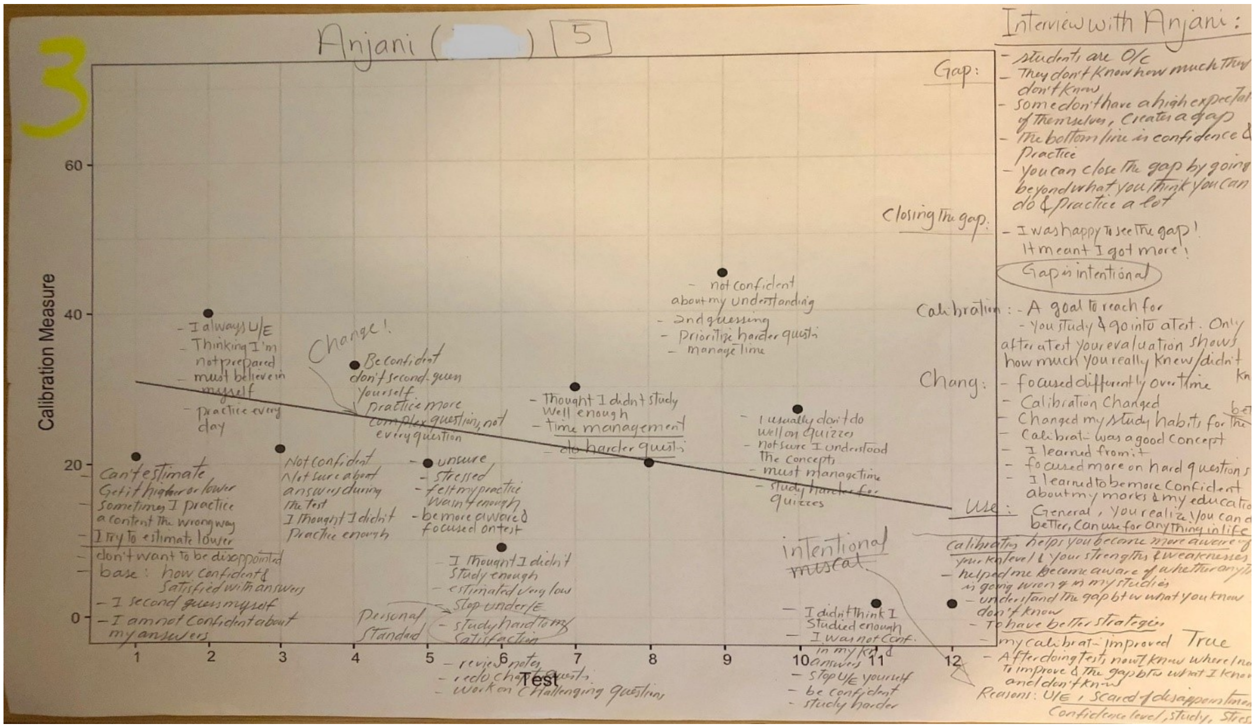
1. Abrar's Blended Data Analysis



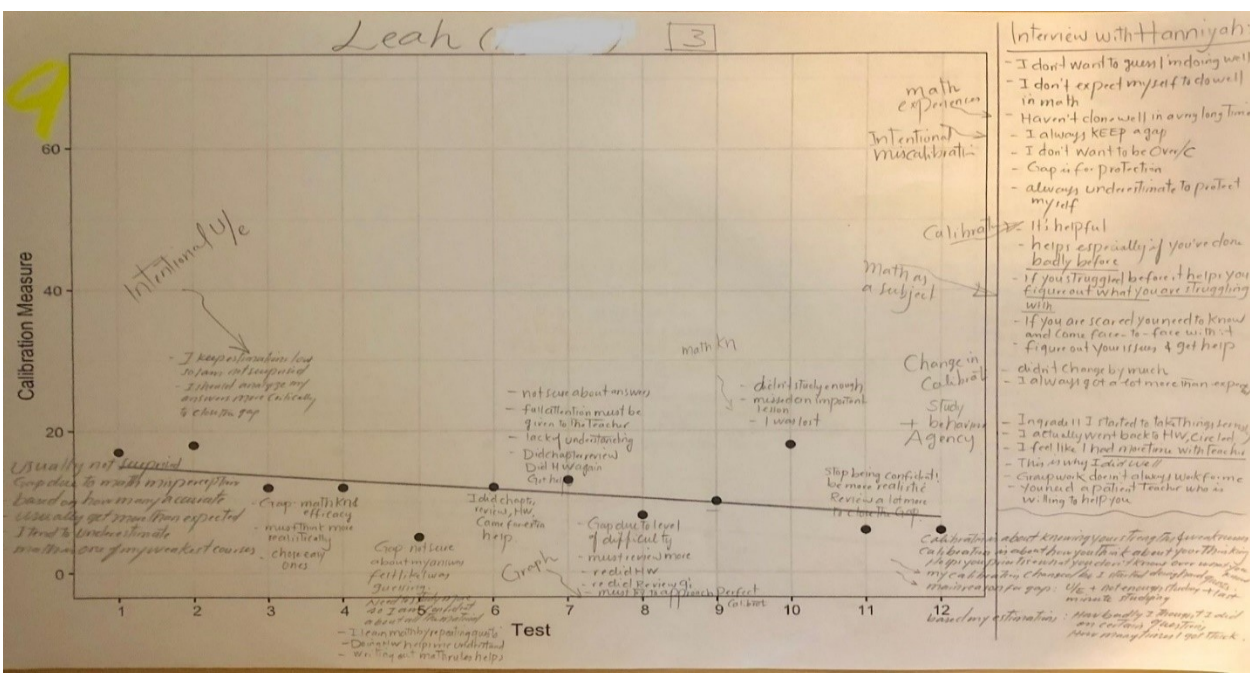
2. Adnan’s Blended Data Analysis



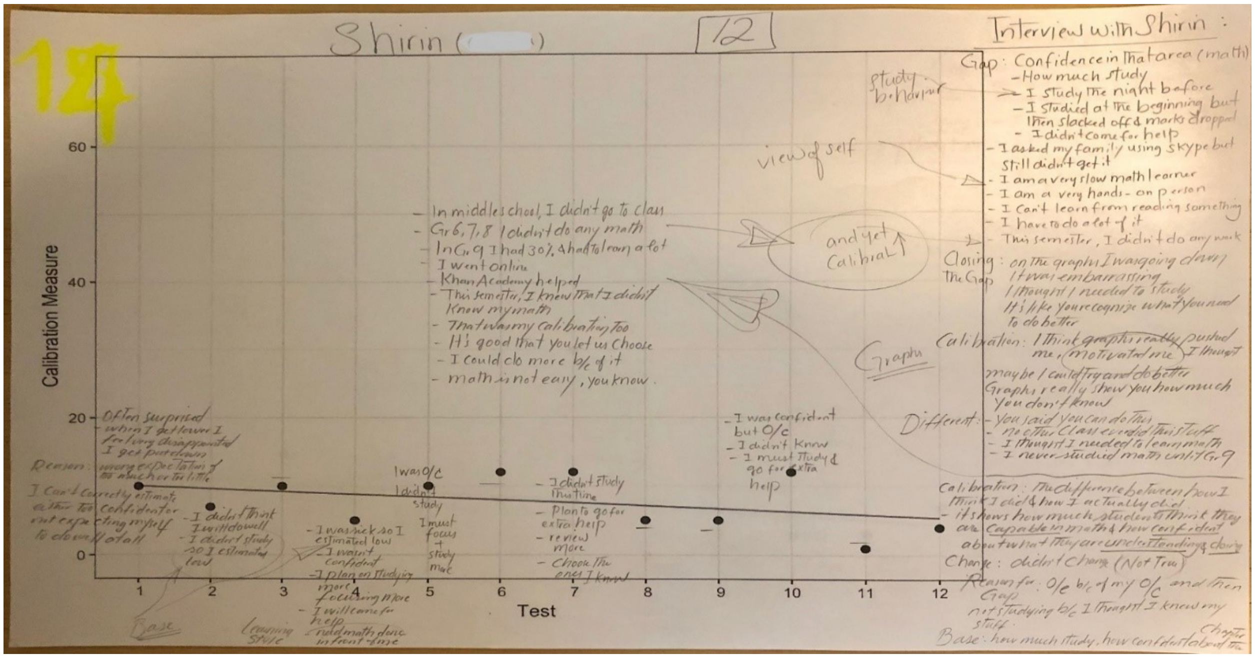
3. Andreas’ Blended Data Analysis



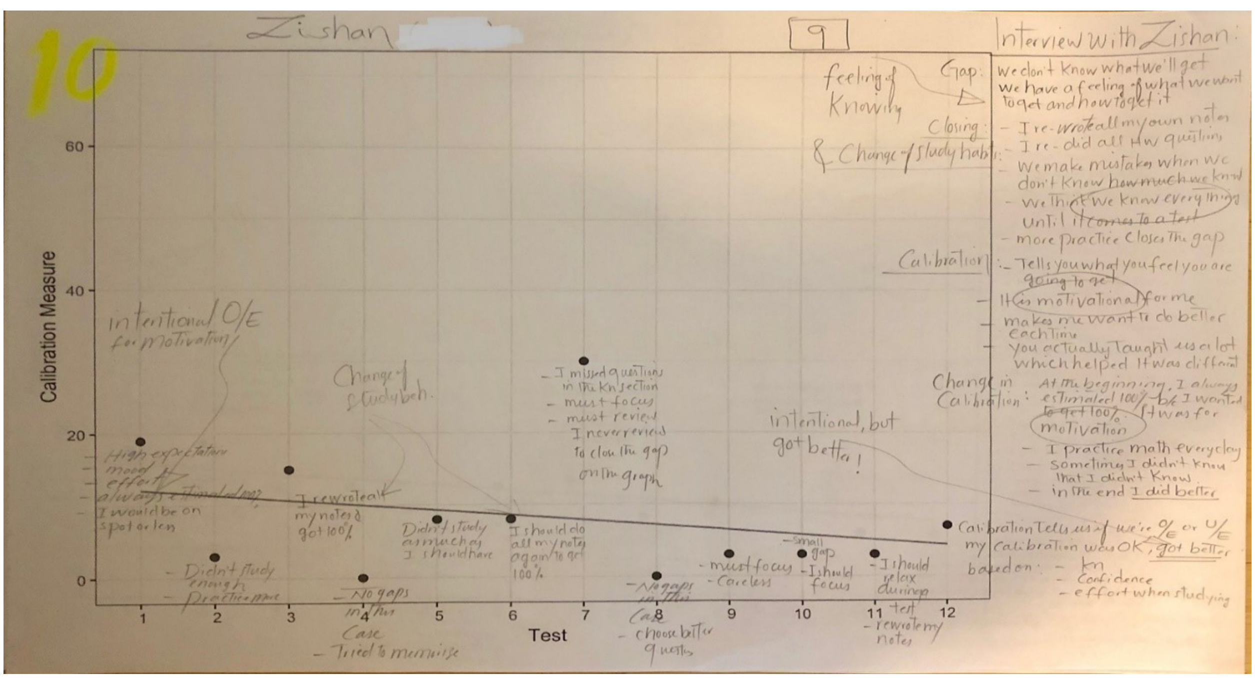
4. Anjani’s Blended Data Analysis



5. Leah’s Blended Data Analysis



6. Shirin's Blended Data Analysis



7. Zishan's Blended Data Analysis

Appendix M

Ethics Approval Issued by the University of Alberta Research Ethics Board (REB 2)

Notification of Approval

Date:	September 28, 2015
Study ID:	Pro00057546
Principal Investigator:	Behnaz Herbst
Study Supervisor:	Florence Glanfield
Study Title:	A Classroom Study of the Development of Mathematics Students' Metacognitive Calibration
Approval Expiry Date:	Tuesday, September 27, 2016

Approved Consent Form:	Approval Date	Approved Document
	9/28/2015	Letterhead_Letter of Invitation (Parents) and Informed Consent Form
	9/28/2015	Letterhead_Research Information Sheet and Contact Information
	9/28/2015	Letterhead_Letter of Invitation (Students) and Assent & Consent Forms

Thank you for submitting the above study to the Research Ethics Board 2. Your application has been reviewed and approved on behalf of the committee.

A renewal report must be submitted next year prior to the expiry of this approval if your study still requires ethics approval. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Approval by the Research Ethics Board does not encompass authorization to access the staff, students, facilities, or resources of local institutions for the purposes of the research.

Sincerely,

Stanley Varnhagen, PhD
Chair, Research Ethics Board 2

Note: This correspondence includes an electronic signature (validation and approval via an online system).

Appendix N

Ethics Approval Issued by the School Board External Research Review Committee (ERRC)

January 12, 2016

Dear Behnaz Herbst,

Re: A Classroom Study of the Development of Mathematics Students' Metacognitive Calibration

Thank you for your detailed response to the External Research Review Committee (ERRC) letter from November 2015 and which addressed all of the comments and questions. On behalf of the ERRC, final approval is now granted for your investigation of students' self-evaluation and meta-cognitive calibration of their task performance in mathematics. As previously noted, we understand that you will be implementing classroom interventions and math tasks to all students as part of the regular program and pedagogy in your own Grade 11 Math class at (School Name), but that only those with consent will submit materials for the research study (e.g. questionnaires, grade estimates, reflection notecards, gap graphing, and follow-up interviews). We also appreciate the change in procedure to conduct individual student interviews on-site before the end of the semester. As a condition of this approval, ERRC will also look forward to receiving both an electronic and paper copy of your final report upon completion, and which you anticipate will be available by …………..

Sincerely,

Sally Erling,

Chair External Research Review Committee,

E-mail: ERRC@....

c.c. Principal, (School Name)

2015-2016-32