Turun yliopisto
University of Turku

INVESTIGATING INDIVIDUAL DIFFERENCES IN AND MEASUREMENT OF ADAPTIVE (RATIONAL) NUMBER KNOWLEDGE IN A SAMPLE OF HIGH SCHOOL STUDENTS

Department of Teacher Education

Master's Thesis

Author:
Irene Pampallis

Supervisor:
Koen Veermans

12.06.2022

Turku

Master's thesis

Subject: Education and Learning
Author: Irene Pampallis
Title: Investigating individual differences in and measurement of adaptive (rational) number knowledge in a sample of high school students
Supervisor: Koen Veermans
Number of pages: 151 pages
Date: 12.06.2022

This thesis examines the performance of a sample of 447 Grade 9-12 students on three different measures of adaptive number knowledge (ANK) with both whole numbers and rational numbers. The aims were to establish whether adaptive whole number knowledge and adaptive rational number knowledge (ARNK) were distinct constructs, to investigate individual differences in A(R)NK in this age group, and to evaluate a new measurement instrument for A(R)NK. The research produced three major findings. First, an exploratory factor analysis of the results on the arithmetic sentence production (ASP) task revealed that adaptive whole number knowledge and adaptive rational number knowledge were not distinct. Second, participant performance on the ASP task was examined using TwoStep cluster analysis and several statistical tests. It was found that the sample performance could be described by a five-cluster model, and membership of high-achieving clusters was strongly associated with mathematics module enrolment and gender, but not with age or school grade. Third, another exploratory factor analysis found that the test comprising of the ASP task and two new tasks was unidimensional, which suggested that all three tasks may measure adaptive number knowledge. The three-task test is therefore a strong candidate for further evaluation and refinement in the quest for a more comprehensive measure of adaptive number knowledge. Each of the three tasks was also examined closely and suggestions for improvements in future iterations of the test were given.

Keywords: Adaptive number knowledge; adaptive rational number knowledge; ANK; ARNK; adaptive expertise; arithmetic; rational numbers.

Acknowledgements

Writing a thesis might seem like an individual project, but it isn't. I would like to express my sincere gratitude to the following people and organisations, without whom this would not have been written:

- Koen Veermans, my supervisor, for his help in clarifying my thoughts and improving my analysis, and his generosity of time and spirit.
- Jake McMullen, for giving me the opportunity to work with this data set and guiding me through the literature and data capture, and suggesting directions for data analysis.
- The Research Group for Mathematics Learning and Instruction at the University of Turku, who gave me useful feedback on my research on several occasions.
- Jo Van Hoof and Hilma Halme, for encouraging me through the final stages of thesis writing, for helping me develop my ideas, and for reminding me that research can be fun and stimulating.
- The Skye Foundation and the FirstRand Foundation, for the funding that made studying abroad possible.
- Robbie Freeman, Magda Czarnecka, Lili Wang and Romy Valo, for their companionship during thesis working sessions at various times during the last two years.
- Kirstin Koivunen, Sirpa Hänti and Mária Kubincova, for our thesis accountability group.
- Gillian Finchilescu, for helping me understand factor analysis and various other statistical conundrums.
- My parents, John and Karin Pampallis, for harassing me to finish this thesis when everyone else was diplomatically avoiding the subject, and for loving and supporting me since the day that I was born.
- My fiancé, Johann Jungbauer, for the pep talks, the dance breaks, making me food, giving me hugs, and generally being a wonderful and supportive partner.

Table of Contents

List of Abbreviations

AIC	Akaike's Information Criterion
ANK	Adaptive number knowledge
AP	Advanced Placement
ARNK	Adaptive rational number knowledge
ASP	Arithmetic sentence production (task)
CR	Correct rejection
EFA	Exploratory factor analysis
FA	False alarm
KMO	Kaiser-Meyer-Olkin (measure of sampling adequacy)
PCA	Principal Components Analysis
SD	Standard deviation(s)
SDT	Signal Detection Theory
US(A)	United States (of America)

List of Figures

List of Tables

1. Introduction

1.1 The Problem of Inflexible Mathematics Learning

It has long been acknowledged that there is a disconnect between mathematics as students learn it in school and mathematics as it is applied to real-world situations. School students have a tendency to learn mathematical procedures in a rote and inflexible manner, with little appreciation of their conceptual meaning or their potential for generalisation. This tendency is recorded in the literature at least as early as 1945. Wertheimer (1945/2020) described a school classroom where the students had learned how to find the area of a parallelogram in which the base was longer than the perpendicular height. They had mastered the procedure, and had proved able to find the area of similarly oriented parallelograms of varying sizes and angles. However, when Wertheimer asked them to find the area of a parallelogram where the base was shorter than the perpendicular height, many students were at a loss. This, despite the fact that the problem would be identical to those the students had already solved if they only rotated the diagram through 45 degrees. Although nearly 70 years have passed since Wertheimer's account was published, the situation he described would be gloomily recognisable to many mathematics teachers today.

Wertheimer's observations of unthinking rote learning are echoed in research from more recent decades. Numerous authors have noted that while students may learn to follow mechanical procedures in school mathematics classes, they often gain little conceptual understanding of what it is that they are doing, and therefore they struggle to transfer their knowledge to unfamiliar test questions or to out-of-school contexts (Boaler, 1998; Graven et al., 2013; Schoenfeld, 1988). Perhaps one of the most striking tales is told by Carraher, Carraher and Schliemann (1985), who investigated the mathematics used by children working as street vendors in Recife, Brazil. The researchers tested the children both informally and formally. Informal tests were conducted verbally at the children's places of work, in the context of making a purchase. Formal tests were conducted as a pencil-and-paper task, containing some context-free mathematical calculations and some word problems. Importantly, the questions asked in the formal tests were based on questions that the children had been able to answer correctly in the informal setting. For instance, in the informal setting, the researcher might first establish that one coconut costs 35 cruzeiros and then ask how much 10 coconuts would cost.

If the child answered correctly, they might then be asked to calculate 35 × 10 in the formal task.

In the informal tests, the children were able to correctly answer 98.2% of questions posed. However, they performed far worse in the formal tests, answering 73.7% of the word problems (which provided some context) correctly, and only 36.8% of the context-free calculations correctly. A qualitative analysis of the children's solution strategies suggested that the errors in the formal tests stemmed from trying to apply school-prescribed algorithms without understanding or reference to number sense, while in their informal calculations they used different (mental) calculation strategies. It was clear that the children saw little connection between the formal mathematical routines taught in schools and the calculations they carried out in daily life. However, and just as importantly, this example also illustrates that school-aged children *are able* to understand and use mathematics – all the street vendors were able to perform the calculations perfectly in their daily working context; they just did not connect these experiences and skills with the mathematics they encountered in the classroom.

Of course, it is not the case that every child fails to grasp the connection between the contents of their mathematics classes and mathematical situations in other subjects or in daily life. Some schools actively endeavour to promote a more critical and flexible understanding of mathematics – for instance, Boaler (1998) described a school which used open-ended projects to structure all mathematics learning – and even in schools which use more traditional methods, some students inevitably develop good number sense and an appreciation for how (school) mathematics can be used in daily life. Nevertheless, the problem is widespread and has long been considered in the literature. For much of the last century, mathematics educators and researchers have been writing about the importance of developing in students a flexible approach towards mathematical problems (Verschaffel et al., 2009). A flexible approach is generally considered to require conceptual understanding as well as procedural skill, and indeed the argument about whether concepts or skills should be weighted more heavily has shaped decades of debate about the optimal method of teaching mathematics in schools (especially in the United States – see Baroody, 2003, for a synopsis of the competing approaches).

1.2 Adaptive Expertise

In recent decades, the distinction between inflexible, rote mathematical skill and more flexible, generalisable mathematical skill has been frequently described in terms of *routine expertise* and *adaptive expertise*. These terms were first introduced by Hatano (1982; also see Hatano & Inagaki, 1984), who initially wrote about routine and adaptive expertise in the contexts of farming, cooking and abacus operation. The terms were later adopted by researchers in various fields related to learning and cognition, and specifically in the fields of mathematics learning and teaching.

Adaptive expertise was defined by Hatano (2003, p. xi) as "the ability to apply meaningfully learned procedures flexibly and creatively". The key to this definition is that procedures are learned *meaningfully*. This means that adaptive experts not only learn how to perform procedures, but they come to understand why, how and when the procedure is effective – and similarly, when and how it should be adjusted (Hatano & Inagaki, 1984). This is what distinguishes adaptive expertise from *routine expertise*: routine experts can apply a procedure quickly, accurately and automatically, as long as they are solving familiar problems in a stable environment. However, they lack the conceptual understanding that allows them to flexibly adapt to new problems or environments (Hatano & Inagaki, 1984). Adaptive expertise thus demands a rich web of connections between various types of conceptual and procedural knowledge (Baroody, 2003; McMullen et al., 2020).

Clearly, it would be preferable to develop adaptive (rather than merely routine) expertise in school mathematics classes. After all, the ultimate aim of teaching students mathematics is not that they are able to pass exams with flying colours. Rather, it is that they are able to understand and use mathematical concepts and techniques to help them flourish in the wider world. They should be able to determine how much it would cost to recarpet their living room, to understand statistics in the newspapers, to scale up a recipe, and to master the higher maths required should they wish to become an architect or an aeronautical engineer or a programmer of computer simulations. The nature of adaptive expertise in basic mathematics – what it is, how to measure it, and how it is developed – is therefore an important topic to understand. This paper aims to contribute to a deeper understanding of adaptive expertise in the field of basic arithmetic, with a specific focus on a subcomponent of adaptive expertise known as adaptive number knowledge (ANK).

1.3 Adaptivity or Flexibility?

Adaptive expertise involves flexibility, but they are not necessarily one and the same thing. The terms *adaptivity* and *flexibility* are used in different ways by different authors, some using them synonymously and others using them to refer to distinct constructs (Heinze et al., 2009). This paper will adopt the distinction drawn by Verschaffel et al. (2009), using the term *flexibility* to refer to the use of multiple strategies, and the term *adaptivity* to refer to selecting appropriate strategies. In other words, adaptivity is about more than simply being able to solve a problem in several different ways – it is about being able to select the most appropriate way (or at least, *an* appropriate way) to solve any particular problem.

The obvious question is: what constitutes an appropriate strategy? Verschaffel et al. (2009) proposed that an adaptive strategy was one that is well-suited (i) to the task at hand, (ii) to the subject performing the task, and (iii) to the context in which the task is performed.

(i) Task characteristics

For any mathematical problem, there will be multiple possible strategies to solve it. Some of these strategies will be simple: they will place relatively little strain on working memory, and their simplicity will make mistakes less likely. Other strategies may be more cognitively complex or more convoluted. All other things being equal (see ii and iii below), a simpler strategy is usually preferable. Precisely which strategy is simpler or more complex often depends on the specific numbers in the problem.

For example, when subtracting numbers where the answer is less than 100, there are three broad categories of strategies: split, jump and varying strategies (Torbeyns et al., 2006). 'Split' strategies involve splitting the two numbers into tens and units and subtracting them separately. For a problem like 47 – 25, a split strategy works well: 40 – 20 = 20 and 7 – 5 = 2, so adding 20 + 2 gives us the answer of 22.

However, if the units in the subtrahend are greater than the units in the minuend, the split strategy becomes more complicated. For instance, if we attempt the same strategy on the problem 62 – 15, we get: 60 – 10 = 50 and 2 – 5 = –3. Of course, it is possible to continue as above and get the correct answer: 50 + (–3) = 47.

However, the negative number –3 makes the calculation considerably more difficult, and may even render it impossible for a child who has not yet learned about negative numbers (or an adult who has forgotten about them, assuming they progressed far enough in their education to learn about negative numbers in the first place). In this case, it might be more appropriate to use a 'jump' strategy. 'Jump' strategies involve using the minuend as a starting point, and then 'jumping' along the number line to subtract the tens and the units of the subtrahend. Thus, to calculate 62 – 15: 62 – 10 = 52 and 52 – 5 = 47.

For certain number pairings, the most appropriate strategy may be neither split nor jump, but one of what Torbeyns et al. (2006, p. 441) call "varying strategies". Varying strategies, as the name suggests, are not a set type, but vary depending on the numbers in the sum. One example is the 'complementary addition' strategy, which can be used to solve subtraction problems where the difference between the two numbers is small. This strategy involves 'adding on' or 'counting on' from the smaller number until one reaches the larger number. For instance, to calculate 53 – 49, it is a very quick mental or think-aloud strategy: "49 … 50, 51, 52, 53", in other words 49 + 4 = 53. And so, 53 – 49 = 4.

All three of the strategies mentioned above *could* be used to solve any of the three problems presented. However, they do not work equally well for each of the three problems (as similarly argued by Torbeyns et al., 2006). The split strategy worked well for the first problem, but gave rise to a bothersome negative value in the second problem. The complementary addition strategy provided a faster solution to the third problem than either of the other strategies would have done, but it would be rather tedious (not to mention accident-prone) to use counting-on to find the solution to the first problem. This illustrates how the most appropriate solution strategy (i.e. an adaptive strategy choice) is dependent on the specific characteristics of the task at hand; the numbers and operations involved determine which strategies are well-suited to the problem.

(ii) Subject characteristics

The chosen strategy must also be an appropriate choice for the person (or subject) performing the strategy. Although a particularly strategy may be well-suited to the

task characteristics of a certain problem, it is only a good choice if the person performing the strategy is able to execute it successfully. For instance, although the jump strategy is a good strategy for subtractions like 62 – 15, it will only work if the subject is comfortable counting down in tens from the number 62. If they are not, a different strategy would be a better choice for them. Similarly, one student may be able to execute the split strategy rapidly, by splitting the tens and units and calculating their differences mentally, while another student may not yet be able to perform the process in their head, and thus will painstakingly write out every step of the process. This would result in the same strategy being faster for the first student than for the second. An adaptive strategy choice requires selecting a strategy that "yields the best performance … in terms of accuracy and speed" (Verschaffel et al., 2009, p. 340), and the best strategy may therefore vary from person to person.

(iii) Context characteristics

An adaptive strategy choice is one that is appropriate to the context in which a problem is solved (Verschaffel et al., 2009). Certain computational tools (e.g. calculators, rulers) may be available in some contexts but not in others. Also, different situations may emphasise different outcomes, such as speed or accuracy. For instance, if one is solving a question in a high-stakes exam, it may be worthwhile to adopt a slightly slower strategy which ensures a perfectly accurate answer. (Alternatively, if the exam is characterised by extreme time constraints, it might be more adaptive to choose the fastest strategy which is still likely to have a high degree of accuracy.) But if one is keeping a running total of costs as one fills up the trolley in a grocery store, it is likely more useful to choose a quick strategy that gives a reasonably close approximation of the total (for instance, by rounding the cost of each item to the nearest 10 cents).

Such contextual factors tend to be obvious and are easily manipulated in experiments, but sociocultural context factors – which operate in the background and thus are more difficult to operationalise and control – have also become increasingly recognised as an important influence on individual strategy choice (Verschaffel et al., 2009). Certain (sub)cultures may place value on, for instance, speed or accuracy, mental or written solutions, unique or conformist solutions, and independent or help-seeking approaches. This can be seen clearly in the example of

> Brazilian street vendors mentioned above (Carraher et al., 1985). In the formal written tests, the children used school-prescribed computational routines, even though they clearly did not understand these routines and their use resulted in less accurate solutions than the mental calculation strategies they used at work. This suggests that they operated under the assumption that in written mathematics, certain (school-taught) algorithms were expected, and that conforming to the expected strategy was more important than obtaining an accurate solution by using a different strategy. Thus, an adaptive strategy choice is also influenced by the characteristics of the (sociocultural) context in which the problem must be solved.

Verschaffel et al.'s definition of adaptive strategy choice – "the conscious or unconscious selection and use of the most appropriate solution strategy on a given mathematical item or problem, for a given individual, in a given sociocultural context" (2009, p. 343) – has been widely adopted (e.g. Brezovszky et al., 2019; Elia et al., 2009; McMullen et al., 2017). However, Selter (2009) notes that this definition understates the importance of creativity, which was central to Hatano's original characterisation of adaptive expertise. Selter therefore suggested that Verschaffel et al.'s definition be modified to explicitly include creativity, as follows: "Adaptivity is the ability to creatively develop or to flexibly select and use an appropriate solution strategy in a (un)conscious way on a given mathematics item or problem, for a given individual, in a given sociocultural context" (p. 624). This paper will follow Selter's definition of adaptivity.

Adaptivity is not a binary variable. Mathematical expertise can be more or less adaptive, and it is also possible that a person could exhibit a sensitivity for some aspects of adaptivity and not others. For instance, Threlfall (2009) gave a hypothetical example of a student who is sensitive to classroom context in deciding on their strategic approach to an arithmetic problem, but who does not consider the number characteristics of the task. Furthermore, each aspect of adaptivity is likely to exist on a spectrum – a person can be more or less responsive to task characteristics, personal characteristics and context.

Threlfall (2009) suggested that strategic flexibility based on the numerical characteristics of a task was desirable primarily because it developed a mathematical way of thinking about numbers and arithmetic operations, which is instrumentally valuable for further mathematical development. He argued that sensitivity to the mathematical characteristics of a task was

therefore uniquely valuable in developing "broad mathematical competence" (p. 543), while sensitivity to task-extrinsic factors was less valuable for this end. Threlfall's argument gives good reasons to place task characteristics front and centre in the conversation about adaptivity with arithmetic, and so it is to a closer examination of task-specific adaptivity that this chapter now turns.

1.4 Adaptive Number Knowledge

1.4.1 Defining Adaptive Number Knowledge

If adaptivity requires selecting an appropriate solution strategy from many possibilities, the next question that arises is: how can someone recognise *which* strategy is the most appropriate for a given arithmetic problem? One answer that has been suggested is that the decision is facilitated by high levels of *adaptive number knowledge*. ANK was defined as "a rich network of knowledge about characteristics of numbers and the [arithmetic] relations between numbers, which can be flexibly applied in solving novel arithmetic tasks" (McMullen et al., 2016, p. 172). In other words, a student with ANK has an awareness of the properties of the numbers in a given problem and a sense of how the numbers might relate to each other. For instance, do the numbers share a common factor? Are they close together? Is one a multiple of the other? Would rearranging the calculation in line with the commutative principle make things easier? And so on (Brezovsky et al., 2019; McMullen et al., 2017). The greater this awareness, the more likely a student will be able to identify an efficient strategy which exploits the relations between key numbers in the problem.

McMullen et al. (2016) gave examples of where children use adaptive number knowledge. One example was the case of additive composition, where the expression 43 – 5 + 7 could be solved more easily if one noticed that 43 + 7 makes a round number. This would require knowledge of number bonds as well as commutativity. Another example was the decision to use the "short-jump" procedure in a subtraction item like 82 – 79, where the numbers in the problem lie on either side of a round number. In the short-jump procedure, the problem solver would first jump from the subtrahend to the round number (79 +1 = 80) and then on to the minuend (80 + 2 = 82), combining both jumps to find the answer (1 + 2 = 3). Thus, it is clear that well-developed ANK could allow children to select more adaptive strategies.

Incidentally, the act of selecting a strategy need not be a conscious decision (Verschaffel et al., 2009), and indeed it may not even be *selected* – in the sense of picking it from a finite list of strategies – at all. Threlfall (2009) suggested that calculation strategies may be constructed afresh for each problem through a process of experimental partial calculations, which build up to a complete correct calculation. In this process, which Threlfall called "zeroing-in" (p. 547), the particular numerical characteristics of the problem inspire the partial calculations, and thus noticing them is essential to the development of a solution. In other words, ANK would be a critical part of the process.

Not all children develop calculation strategies in situ like this; some are more likely to rely on pre-learned algorithms (McMullen et al., 2017). However, even in the scenario where children rely on pre-learned algorithms, choosing the most efficient strategy from among all the strategies one has learned depends on recognising a strategy that is well-suited to the problem at hand, something that often requires an awareness of the relation between the numbers in the problem – in other words, it requires ANK.

ANK is theorised to work in tandem with procedural flexibility: procedural flexibility means that someone knows *how* to use a range of calculation strategies; ANK helps them choose *which* strategy to use. Both are necessary for adaptive expertise with arithmetic, and neither alone is sufficient to explain adaptive expertise (McMullen et al., 2016). Both these elements – ANK and procedural flexibility – are regarded as behavioural manifestations of adaptive expertise (McMullen et al., 2022).

1.4.2 Measuring Adaptive Number Knowledge

ANK is measured with an instrument known as the arithmetic sentence production (ASP) task, developed by McMullen et al. (2016). In this time-limited task, research subjects are presented with a set of five given numbers and a target number. Subjects are required to generate as many arithmetic sentences as possible, using some or all of the given numbers, which equal the target number. The four basic arithmetic operations (addition, subtraction, multiplication, division) may be used. In order to generate a large number of solutions, subjects must rapidly notice the characteristics of the given numbers and potential arithmetic connections between them.

Most past research on whole number ANK has made a distinction between *dense items* and *sparse items* in the ASP task (McMullen et al., 2016, 2017, 2019). In dense items, there are many easily identifiable connections between the given numbers and the target number. For instance, the numbers will have several common factors and multiples, and there will be many single-step procedures that lead to the target number. The target number will typically be thoroughly familiar to the research subjects, the sort of number that comes up often in classroom or daily calculations. This means that most people will find it fairly easy to generate a large number of solutions. By contrast, in sparse items there are relatively few obvious relations between the given numbers and the target number. Multi-step procedures are usually necessary to reach the target number, and the research subjects should be expected not to have memorised many arithmetic facts about the target number. As a result, it is more difficult to generate many solutions for sparse items, particularly without using multiple operations. Since generating complex solutions demands a more integrated understanding of the potential relationships between numbers and operations, it may signal a higher level of ANK. Examples of one dense and one sparse item from McMullen et al. (2016) are given in Table 1.1.

Table 1.1

Example items from the arithmetic sentence production task in McMullen et al., 2016

Item type	Given numbers	Target number
Dense	2, 4, 8, 12, 32	16
Sparse	1, 2, 3, 5, 30	59

From the outset, research on adaptive number knowledge has made frequent use of person-centred analytical approaches (see for instance McMullen et al., 2016, 2017, 2019, 2020). A person-centred approach can be contrasted to a variable-centred approach, although the two approaches are complementary rather than incompatible (Niemivirta et al., 2019). A variable-centred approach takes variables as the unit of analysis, and tests the relationship between different variables through statistical techniques like correlations and regressions (Niemivirta et al., 2019). The underlying assumption is that variation within the population on each variable is purely quantitative, not qualitative; in other words, the relation between variables is the same for everyone in the sample (Hickendorff et al., 2018). However, such an approach may obscure the existence of distinct groupings within the population, which differ systematically on the variables in question. A person-centred approach, by contrast, can reveal heterogeneous patterns of variables which might otherwise be overlooked. This is particularly important in

educational contexts, where the "average" learning pattern may not describe many (or any) subgroups of learners adequately (Hickendorff et al., 2018).

The first studies on ANK (reported in McMullen et al., 2016) used K-means cluster analysis to identify subgroupings of participants based on their performance on the ASP task. Later studies (McMullen et al., 2017, 2019, 2020) have made use of latent variable models, a group of more sophisticated techniques which make fewer assumptions about variables, are more flexible, and can handle longitudinal data (Hickendorff et al., 2018; McMullen et al., 2017). This approach has revealed interesting patterns of whole number ANK among students in upper primary school and lower secondary school (see Section 1.4.3).

The fact that only one task is used to measure ANK has been identified as a limitation of the existing research (McMullen et al., 2019, 2022). Relying on a single measure, even if it is highly reliable, means that it is impossible to tell to what extent it captures the underlying construct, as opposed to measurement error that is specific to the particular task. It may also capture only a limited portion of the construct we call "adaptive number knowledge". A wider range of assessment types is needed to determine the precise nature of ANK (McMullen et al., 2019) and to improve its interpretability (McMullen et al., 2020).

1.4.3 Previous Studies of Whole Number ANK

Only two studies on this topic have used person-centred analysis with large samples, one with Finnish Grade 4-6 learners (McMullen et al., 2017, n = 1065) and one with Finnish Grade 7 learners (McMullen et al., 2019, n = 879). McMullen et al. (2017) conducted a latent profile analysis based on how many correct responses each student produced for dense and sparse items, and on whether the responses were *simple* (i.e. using only additive or only multiplicative operations) or *complex* (i.e. using both additive and multiplicative operations). They identified five profiles. Just over 50% of students were in the Basic profile, which produced a below-average number of correct responses for all item and response types. A very small proportion of the students (2.3%) were in the High profile, which produced a relatively large number of correct responses for all item and response types. The remaining students were split between three profiles, labelled Simple (5.1%), Complex (14.9%) and Strategic (27.5%). Learners in these profiles produced a similar number of correct responses, but the nature of their solutions

differed between groups. The Simple profile produced almost exclusively simple responses on both dense and sparse items. The Complex profile was characterised by producing an above-average number of complex responses on both dense and sparse items, although they also produced many simple responses on the dense items. The Strategic profile adapted its approach depending on the item type, producing mainly simple responses on the dense items and mainly complex responses on the sparse items.

A similar profile structure was found in the confirmatory latent profile analysis run by McMullen et al. (2019). The Basic profile was again by far the largest (58.8% of learners) and the high profile the smallest (4.0%). The Simple (10.2%), Complex (19.6%) and Strategic (7.4%) profiles performed similarly to the 2017 study, but represented different proportions of the total sample.

These findings are interesting for several reasons. Firstly, they illustrate that the majority of children seem to fall into the Basic profile, demonstrating that even among children who have completed the majority of their primary school education, in a country with a good standard of education, the ability to rapidly and flexibly calculate answers to a novel whole number arithmetic problem is highly constrained. (It should be noted that learners in the Basic profile still produce a non-trivial number of solutions: an average of 8.1 and 12.32 correct solutions across all four items in the 2017 and 2019 studies respectively. However, this is far lower than the next-best profiles' totals of 13.54 and 20.26.)[1] This relatively low level of ANK helps to explain the general lack of adaptive expertise in arithmetic that was highlighted at the beginning of this chapter. Another implication of this distribution is that differentiation between different types and degrees of ANK largely happens at the upper end of the mathematical skill spectrum, as suggested by McMullen et al. (2019). Secondly, these findings reveal that even among students with similar overall scores on the ASP task, there can be substantial variation in their overall strategic approach.

It should be noted that ANK is a relatively new object of study, and what defines "high" ANK is still a live question. Some research has measured ANK purely quantitatively, with a higher score on the ASP task taken to reflect a higher level of ANK (Brezovszky et al., 2019; Kärki et al., 2021; McMullen et al., 2022). Other research has combined quantitative measurements

[1] These totals are from my own calculations, based on figures given in the relevant articles.

with a qualitative person-centred analysis of solution types to create a ranking based on both, as in the Basic-Simple-Complex-Strategic-High hierarchy mentioned above (McMullen et al., 2016, 2017, 2019). This ranking does not correspond perfectly with a purely quantitative approach: in terms of total correct responses, the Simple profile actually produced slightly more responses than the Complex profile (McMullen et al., 2017, 2019) and in one case also more than the Strategic profile (McMullen, 2017), but because of the uniformly simple nature of its responses, it was regarded as lower on the ladder of ANK. The inter-profile differences on measures of arithmetic fluency and conceptual knowledge noted above were also taken to justify this position in the hierarchy.

Importantly, ANK research to date has focused almost exclusively on younger students (Grade 3-8). There have been no large-scale studies of older students. As a result, we lack a clear account of the differential developmental trajectories of ANK as students progress through high school and into higher education or the workplace. This is a major limitation of the status quo. Research on older students would be able to address questions such as whether ANK development levels off after a certain age and whether all students eventually reach the same level of ANK. If they do not, this may help to explain why students have differing rates of success in higher mathematics, including the transition to algebra.

1.4.4 Correlates of ANK

ANK profile membership has been found to be related to several other mathematical variables, most importantly arithmetic fluency. McMullen et al. (2017) found strong relationships between profile membership, arithmetic fluency and arithmetic conceptual knowledge, with arithmetic fluency and conceptual knowledge increasing significantly from profile to profile in the order: Basic, Simple, Complex, Strategic, High. Arithmetic fluency had a particularly large effect on profile membership. As the authors pointed out, this is unsurprising on a time-limited task, but it is telling that there are fluency differences even between the three middle profiles, which have similar overall scores, suggesting that the ability to calculate rapidly supports the creation of more complex or adaptive solutions.[2] In McMullen et al. (2019), a similar

[2] Some authors have speculated that these differences could be due to differences in working memory, since the ASP task requires that the target number be kept in mind while working forwards to or backwards from the desired solution (McMullen et al., 2019). This could place an increasingly large burden on working memory as the potential solution becomes more complex, with more intermediate values. Individuals with lower working

relationship was found between conceptual knowledge, arithmetic fluency and profile membership, although the differences between the Complex, Strategic and High profiles were not significant. This further supports the idea that ANK differentiates between children who already have a high level of mathematical knowledge and skill (McMullen et al., 2019). ANK profile membership has also been shown to be a unique predictor of pre-algebra skills (McMullen et al., 2017).

1.5 Rational Numbers and Adaptive Number Knowledge

Rational numbers are numbers which can be written in the form $\frac{a}{b}$ where both a and b are integers and b is non-zero. They may be represented in a number of ways, most commonly as fractions, terminating decimals or recurring decimals, but also in several other forms, including percentages and ratios. An understanding of rational numbers and rational number operations is essential for many types of higher mathematics, like algebra, geometry and statistics, as well as many mathematics-adjacent fields and careers, and even careers that are not normally thought of as mathematical, such as nursing (Bailey et al., 2014; Lortie-Forgues et al., 2015). In school children, rational number understanding has been found to predict algebra readiness and general mathematical achievement (Van Hoof et al., 2017), as well as algebra knowledge and mathematics achievement (Bailey et al., 2014).

However, rational numbers are also a topic that many children and adults struggle with (Bailey et al., 2014; Kärki et al., 2022; Siegler & Pyke, 2013; Van Hoof et al., 2017). Fraction difficulties tend to compound with time: early difficulty with fractions predicts later difficulty with fractions (Hecht & Vagi, 2010; Mazzocco & Devlin, 2008), and Siegler and Pyke (2013) found that the gap in fraction arithmetic knowledge between low-achievers and their classmates increases between Grade 6 and Grade 8, with the stronger students improving significantly while their lower-performing peers stagnated.

memory capacity may therefore struggle to use more complex solution strategies (Threlfall, 2009). However, the only study (to the author's knowledge) that has examined the relationship between working memory and performance on the ASP task found no significant relation between profile membership and working memory (McMullen et al., 2019).

There are many reasons that learners struggle with rational numbers, some of which have been outlined by Lortie-Forgues et al. (2015). In addition to culturally contingent factors, which vary from country to country (relating to factors such as pedagogy, teaching expertise, curriculum and textbook approaches), they identify seven intrinsic sources of difficulty in fraction and decimal arithmetic: (1) mastering new notations, which function in different ways to whole number notation; (2) grasping the magnitude of fractions requires deriving a magnitude from a ratio of two numbers, which is considerably more complex than the process for whole numbers; (3) the conceptual basis of rational arithmetic procedures is often not immediately apparent; (4) there are complex relations between rational arithmetic procedures and whole number arithmetic procedures, with some features of whole number arithmetic carrying over to rational numbers and others not; (5) there are complex relations between the various rational arithmetic procedures; (6) the effects of multiplying and dividing fractions and decimals between 0 and 1 are opposite to what would be expected with whole number arithmetic, but the same as with whole number arithmetic once the absolute value of the rational number exceeds 1; (7) the sheer number of distinct procedures that must be learned and applied, particularly when working with fractions (e.g. whole number arithmetic procedures, finding equivalent fractions, simplifying fractions, converting between mixed numbers and improper fractions, and fractional arithmetic procedures).

A common theme running through many of these difficulties is that rational numbers work differently to whole numbers. The same rules and intuitions that work with whole numbers do not necessarily work with rational numbers. Therefore, it follows that adaptive rational number knowledge (ARNK) would require different (or additional) skills to whole number ANK. While the basic idea – understanding the characteristics of numbers and the connections between them – remains the same, different conceptual and procedural knowledge would be required if the numbers in question are rational numbers. Kärki et al. (2022) suggested that the skills of finding equivalent fractional forms and simplifying fractions, as well as conceptual understanding of rational number magnitudes, rational number arithmetic procedures, and how the base-ten structure can be extended to decimal numbers, are among the skills and procedures which need to be integrated to support ARNK. A critical element of any type of mathematical problem solving is being able to move between different representations (Heinze et al., 2009), and this is particularly important with rational numbers. In order to be able to solve rational number problems adaptively, representational knowledge is essential: for instance, one must

know that the fraction $\frac{1}{4}$ can also be expressed in decimal form as 0.25, or $\frac{3}{12}$, or any one of a host of other equivalent fractions (Kärki et al., 2022; McMullen et al., 2020).

ARNK has never been compared to (whole number) ANK. It is therefore unclear how closely related the two constructs are. On the one hand, adaptivity with rational numbers demands a host of skills and knowledge that are not required for adaptivity in whole number problem solving – and which many people never master. Therefore, it is plausible that some learners may have well-developed ANK but poor levels of ARNK. On the other hand, some studies have shown connections between rational number achievement and whole number achievement. Rinne et al. (2017) reported that general mathematics achievement and whole number line estimation ability predict both performance and improvement on a fraction magnitude comparison task; Bailey et al. (2014) found that knowledge of whole number arithmetic and whole number magnitudes in Grade 1 predicts fraction magnitude knowledge and fraction arithmetic knowledge in middle school; and Siegler and Pyke (2013) observed that whole number division skills predict performance on a fraction arithmetic task. This might suggest that the same students who display high levels of ANK (and who therefore have strong and well-integrated whole number knowledge) are likely to go on to develop the rational number knowledge on which well-developed ARNK is predicated. The question of whether ANK and ARNK are in fact distinct is one of the questions that this thesis aims to answer. A clear answer to this question will contribute to a clearer characterisation of adaptive number knowledge, and may also contribute to a better understanding of the interrelationship between whole and rational number understanding.

Although ANK and ARNK have never been directly compared, a few studies focusing on adaptivity with rational numbers give insight into the nature of ARNK specifically. (ARNK is measured similarly to ANK, with the arithmetic sentence production task. The only difference is that some of the numbers in the task are fractions and decimals.) Although ARNK is correlated with routine procedural and conceptual knowledge of rational numbers, it is distinct from both of them. McMullen et al.'s (2020) study of 394 American seventh- and eighth-graders found that ARNK is best modelled separately from routine conceptual and procedural rational number knowledge. The authors subsequently conducted a latent profile analysis, the results of which suggested that routine procedural rational number knowledge is a prerequisite for routine conceptual rational number knowledge, and both are prerequisites for high levels of

ARNK. This provides support for the position that well-developed ARNK is a characteristic of high-achieving mathematics students who have already mastered routine procedures and knowledge. The fact that some students display both routine procedural and conceptual knowledge without high levels of ARNK also speaks to the fact that adaptive number knowledge requires more than siloed concepts; integrating concepts with each other and with procedural knowledge is required. The authors make the interesting observation that representational flexibility – in this case, combining decimals and fractions in the same expression, or generating several responses that were mathematically identical but notationally different (e.g. ½ + ½ , 0.5 + 0.5 and ½ + 0.5) – was associated with membership of more adaptive profiles.

The same study found that ARNK was a unique predictor of performance on state-wide standardised algebra tests, underscoring the importance of ARNK for mathematical development more broadly. Another study (McMullen et al., 2022) investigated the predictors of ARNK in a sample of 173 Grade 6 and Grade 7 learners from the USA, and found that 47% of variation in ARNK scores was explained by mathematics-specific predictors (interest, achievement and multiplicative reasoning), rational number magnitude and operations knowledge, and spontaneous focusing on multiplicative relations. Interestingly, domain-general variables such as non-verbal intelligence and spatial reasoning did not have any unique explanatory power.

1.6 The Development of A(R)NK Over Time

Relatively little is known about how A(R)NK develops as children progress through school. The only two studies to have measured A(R)NK at multiple time points (Brezovszky et al., 2019; Kärki et al., 2022) were intervention studies testing the effects of mathematical computer games, and took place over relatively short time spans (2 weeks for Kärki et al. and 10 weeks for Brezovszky et al.). Therefore, these studies cannot contribute to an understanding of the "natural" development of A(R)NK, or the development of A(R)NK over longer periods of time. However, existing cross-sectional research does give some insight into the possible development of A(R)NK over time.

Multi-grade studies suggest that whole number ANK improves with age. McMullen et al.'s (2016) study of 55 students in Grades 3-5 found that membership of more adaptive clusters was associated with school grade, with Grade 3s most likely to be placed in the Basic and Simple clusters, and Grade 5s more likely to be placed in better-performing clusters. A larger study with 1065 Grade 4-6 students (McMullen et al., 2017) found similar results, with students from lower grades more likely to be placed in the Basic profile and less likely to be in the Strategic or Complex profiles.[3] Single-grade studies using similar versions of the ASP task provide further evidence to support this position. The mean number of correct responses in the ASP task seems to increase with the grade of the tested students. McMullen et al.'s (2016) small sample of Grade 3-5 students produced an average of 7.44 correct solutions across four ASP items (own calculations based on figures in the article), while a Grade 4-6 sample (McMullen et al., 2017) produced 10.99 correct solutions on average (own calculations). The trend continued with a sample of 879 Grade 7 students, which produced an average of 16.24 correct solutions (McMullen et al., 2019; own calculations). It is interesting to observe that even the Basic profile in the Grade 7 sample outperformed the mean scores from the earlier studies, producing an average of 12.32 correct solutions. This is even more interesting given the authors' finding that the Basic profile was substantially larger in the Grade 7 sample than in the earlier Grade 4-6 sample. This suggests that, even as the proportion of students who have high ANK relative to their classmates decreases, the overall level of ANK increases. A final observation is that the development of ANK may be related to arithmetic fluency: McMullen et al. (2019) note that students in the Complex, Strategic and High profiles had improved in arithmetic fluency from Grade 6 to Grade 7, while the arithmetic fluency of students in the Simple and Basic profiles had stagnated over the same time period. They suggest that fluency improvements could reflect increasing automatization of basic numerical facts, and that this might support more complex and creative reasoning processes, such as are required by the ASP task.

A relationship between school grade and adaptive rational number knowledge was, however, not found in the only study of ARNK to examine multiple grades (McMullen et al., 2020). In this study of 384 Grade 7 and 8 learners, no connection was detected between profile membership and grade. It is certainly plausible that A(R)NK development levels off at some

[3] It should be noted that these patterns notwithstanding, there was also significant within-grade variation in these studies.

point, but it is surprising that ARNK would already have levelled off in this age group, since Van Hoof et al. (2017) note that rational number understanding develops particularly rapidly between Grades 6 and 8, and is only expected to stagnate in the last years of high school. It is therefore unclear whether ARNK really has levelled off by this point, or whether this result was due to particular characteristics of the sample. More studies are required to answer this question with any degree of certainty.

As noted in Section 1.4.3, no research has been done on students above Grade 8, which also hampers a thorough understanding of A(R)NK's developmental trajectories. Studies with older students could provide a comparison to the existing studies of younger children, and interrogate how and whether A(R)NK continues to develop and be useful even after formal instruction on whole and rational numbers has ended.

1.7 Research Questions

Previous research has investigated ANK and ARNK separately, but no research to date has directly compared the two types of knowledge. It is therefore unknown whether they are distinct constructs or different facets of the same construct. Thus, the first question this paper aims to answer is: ***Are adaptive whole number knowledge and adaptive rational number knowledge empirically distinct constructs?***

Secondly, previous research has focused on children in elementary school and middle school/lower secondary school (depending on the country). Examining the nature of A(R)NK in older students may generate new insights into how A(R)NK develops over time. Therefore, the second question this paper poses is: ***Are there quantitative and/or qualitative individual differences in high school students' ANK and ARNK?*** A sub-question of particular interest is: *Are there systematic differences between students in terms of age or school grade?*

Finally, A(R)NK has so far been evaluated only by means of the arithmetic sentence production task. This has been identified as a limitation of the existing research as it restricts the interpretability of the A(R)NK construct. Thus, the third question this paper aims to answer is: ***Does an instrument consisting of three different tasks provide a more nuanced understanding of A(R)NK than the ASP task provides alone?***

The remainder of this thesis will proceed as follows. Chapter 2 describes the sample and research methods. Chapters 3, 4 and 5 present analysis and discussion aimed at answering the three research questions respectively. Finally, Chapter 6 presents some suggestions for further research and concludes this thesis.

2. Research methods

2.1 Participants

The sample for this study consisted of 447 students (45.2% female, 52.8% male, 0.2% non-binary and 3.8% missing gender data) from a single high school in the south-eastern United States. The school in question was a large public high school with over 2 300 students (53% Caucasian, 21% Black, 12% Hispanic, 7% Asian, 6% multiracial, 1% other). The school was located in a suburban area, and approximately 30% of students qualified for free or reduced-price lunch.

Students were in Grades 9 to 12 at the time of participation, and ranged from 13 to 18 years old. A large majority of the participants (78.3%) were in Grade 9 or 10, and accordingly almost two-thirds were aged 15 years or younger. Complete grade and age distributions are shown in Figures 2.1 and 2.2, with summary statistics in Table 2.1. Participating students were spread across 19 different mathematics classes taught by a total of four teachers. All students agreed to participate in the research; their parents also gave permission. Students were allowed to opt out at any point before or during the testing process. The study was approved by the ethics board of the University of Turku, as well as the district and school administration. The data was collected in January 2020.

Figure 2.1

Bar Graphs Showing the Distribution of Participants by School Grade (left) and by Age (right)

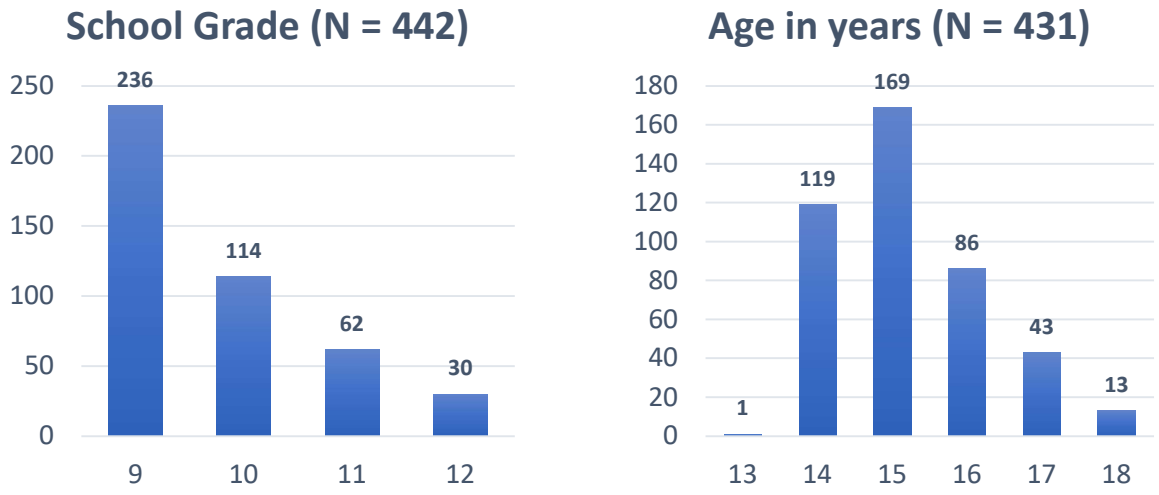


Table 2.1

Descriptive Statistics for Participants' Age and Grade

	N	Mean	Standard Deviation	Skewness	Kurtosis	Range
Age (years)	431	15.21	1.06	0.70	-0.09	13-18
Grade	442	9.74	0.94	1.03	-0.04	9-12

The participants' school, like many high schools in the United States, follows a modular mathematics curriculum. Each year, students register for one or more mathematics courses on specific topics. Course enrolment is not dependent on a student's school grade, but rather on the mathematics courses they have completed previously. High-achieving students often complete some high-school level mathematics courses whilst in middle school, which allows them to register for more advanced classes in ninth grade, the first year of high school. Thus, students of different ages and school grades may be in the same mathematics class.

It is reasonable to suspect that a student's level of mathematical education may have some influence on their performance, so the participants' current mathematics course enrolments were recorded (Advanced Placement [AP] Statistics, Algebra 1 Honors, Algebra 2, Algebra 2 Honors, Geometry or Geometry Honors)[4]. The grades in which each course is offered, and the prerequisites for enrolment, are summarised in Table 2.2. Note that students take *either* Geometry *or* Geometry Honors, and *either* Algebra 2 *or* Algebra 2 Honors. The honors courses are more demanding modules intended for high-achieving students, and replace the basic courses for those topics. Course enrolment is detailed in Table 2.3. Participants' most recent mathematics grade was also recorded. The percentage of students receiving each grade was as follows: A (38.9%), B (31.3%), C (16.8%), D (5.8%), F (1.8%), missing data (5.4%).

2.2 Testing Procedure

The data used in this thesis comes from the first half of a two-part test, in which the first part focused on ANK and the second part focused on flexibility in arithmetic and algebraic problem

[4] Other mathematics courses were also offered by the participants' school, but they are not relevant here since all students in the sample were enrolled in one of the six listed courses.

Table 2.2

Prerequisites for Enrolment in Mathematics Courses

Course name	**Grades in which course may be taken**	**Prerequisites**
Algebra 1 Honors	9, 10, 11, 12	Teacher recommendation
Algebra 2	9, 10, 11, 12	C or above in Algebra 1
Algebra 2 Honors	9, 10, 11, 12	B or above in Algebra 1 Honors *OR* teacher recommendation
Geometry	10, 11, 12	C or above in Algebra 2 *OR* studied concurrently with Algebra 2, with a B or above in Algebra 1
Geometry Honors	9, 10, 11, 12	C or above in Algebra 2 Honors *OR* studied concurrently with Algebra 2 Honors, with a B or above in Algebra 1 Honors + teacher recommendation
AP Statistics	10, 11, 12	Algebra II Honors or Geometry

Table 2.3

Mathematics Course Enrolment for All Participants (N=442)

	Student's school grade				
	9	**10**	**11**	**12**	**Total**
Algebra 1 Honors	42 (18%)	0 (0%)	0 (0%)	0 (0%)	**42** **(9%)**
Algebra 2	0 (0%)	26 (23%)	19 (31%)	8 (27%)	**53** **(13%)**
Algebra 2 Honors	24 (10%)	76 (66%)	29 (47%)	1 (3%)	**130** **(29%)**
AP Statistics	0 (0%)	2 (2%)	7 (11%)	20 (67%)	**29** **(7%)**
Geometry	3 (1%)	10 (9%)	7 (11%)	1 (3%)	**21** **(5%)**
Geometry Honors	167 (71%)	0 (0%)	0 (0%)	0 (0%)	**167** **(37%)**
Total	**236** **(100%)**	**114** **(100%)**	**62** **(100%)**	**30** **(100%)**	**442** **(100%)**

solving. Participants completed a pencil-and-paper test in their mathematics classrooms during normal school hours. The ANK section of the test comprised four components: an arithmetic sentence production (ASP) task, a self-evaluation task based on the ASP task, a missing symbol task, and a rapid verification task. The present thesis does not use data from the self-evaluation test, so it is omitted from the descriptions below. All tasks were introduced and supervised by a researcher.

The intention of the multi-component ANK test was twofold: to investigate the adaptive number knowledge of a sample of high school students; and to trial new measures for testing ANK, which could supplement the established ASP task. The three tasks analysed in this thesis proceeded from a very open-ended, creative task (the ASP task), through a semi-open task which required some creativity but constrained the possible answers much more (the Missing Symbol task), to a completely closed task (the Rapid Verification task). All were expected to activate ANK in certain ways, as outlined in the description of each task below.

2.2.1 Arithmetic Sentence Production Task

The Arithmetic Sentence Production task was included as a sort of "anchor" for the test, as it is a well-established measure of adaptive number knowledge (see Section 1.4.2). This task included one practice item and four test items. For each item, participants were given 90 seconds to generate as many arithmetic expressions as they could, that equalled a given target number. In their expressions, participants were allowed to use any or all of a given set of five numbers and the four basic arithmetic operations (addition, subtraction, multiplication and division). Using a number or operation more than once was allowed. Brackets were also included among the given operations, to make it clear that participants could manipulate the order of operations. An example item is shown in Figure 2.2.

First, the participants completed a practice item with small, familiar numbers (make 6 by combining a subset of 1, 2, 3 and 4). After completing the practice item, they were invited to ask questions about the task. After this, they completed the four test items. Two of the test items contained whole numbers and the other two test items contained rational numbers. In the rational number items, the given sets of numbers consisted of two pairs of equivalent numbers

in fraction and decimal form (e.g. $\frac{1}{4}$ and 0.25; $\frac{3}{4}$ and 0.75) and one whole number. See Table 2.4 for a summary of test items.

Figure 2.2

Example Item from the Arithmetic Sentence Production Task

The instructions read: "Try to make as many different math problems where the solution is 6 as you can. Use only the numbers in the box. You can use each number as many times as you want. You can use addition, subtraction, multiplication, and division as many times as you want. Write your answers in the blank box."

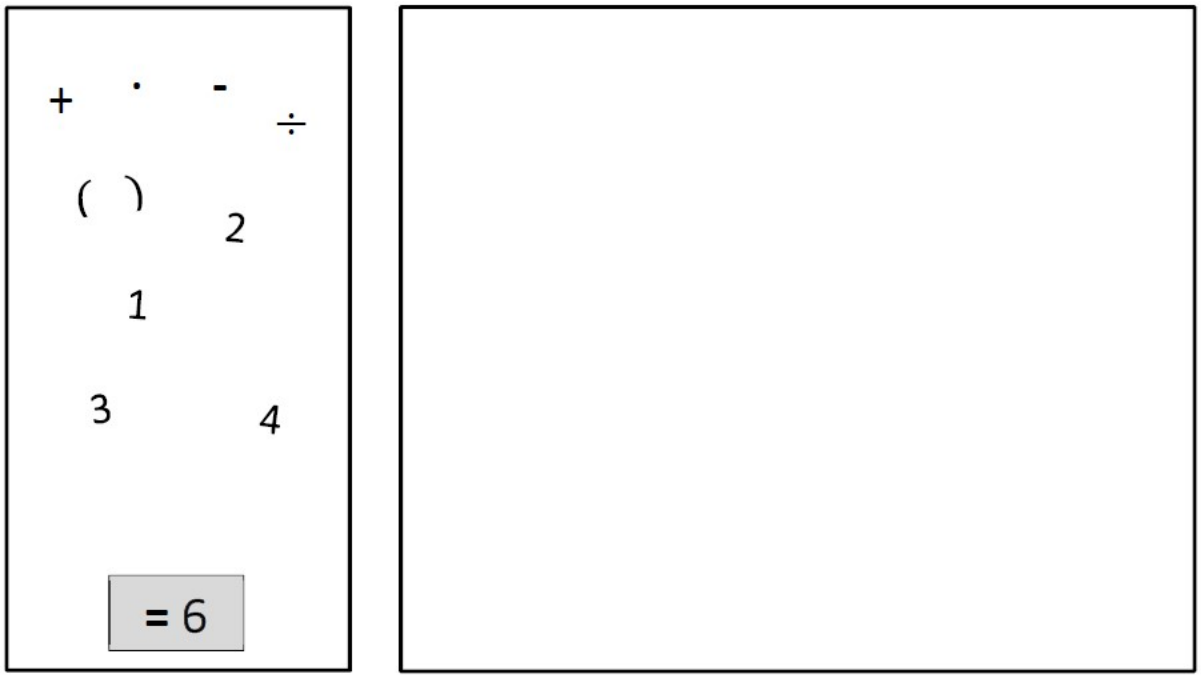


Table 2.4

Summary of ASP Task Test Items

Item	Given numbers	Target number	Problem type
1	1, 2, 3, 5, 30	59	Whole
2	2, 3, 6, 10, 18	38	Whole
3	$\frac{1}{4}$, 0.25, $\frac{3}{4}$, 0.75, 2	$\frac{1}{2}$	Rational
4	$\frac{3}{4}$, 0.75, $\frac{3}{2}$, 1.5, 2	3	Rational

As outlined in Section 1.4.2, most past research has made a distinction between dense items (in which there are many easily identifiable connections between given numbers, and many

single-step solutions that lead to the target number) and sparse items (in which there are fewer connections between given numbers, the target numbers are less familiar, and most solutions require multiple steps). In the present test, the whole number items were sparse items that had been used in previous research (McMullen et al., 2017, 2019). The rational number items had been previously used by McMullen et al. (2020), but had not been classified as dense or sparse. Although they lacked the overwhelming familiarity and volume of obvious solutions as would be found in the dense whole number items from prior research (e.g. make 16 using 2, 4, 8, 12 and 32), the rational number items still had very familiar target numbers (½ and 3) and a substantial number of one-step solutions were easily spottable. For this reason, they will be referred to as dense items.

This means that the whole number items are exclusively sparse and the rational number items are exclusively dense. The decision was made to use only sparse items in the whole number tasks because the participants were in high school and were expected to be able to handle dense items with relative ease, meaning that such items might have diminished power to discriminate between students with higher and lower levels of ANK (since all would be likely to score highly). In the case of the rational number items, sparse items were trialled in a pilot study but participants found them so difficult that they produced barely any solutions, which similarly led to a reduction in discriminatory power. Therefore, denser items were selected for the rational number tasks.

Answers were coded as correct if they were mathematically correct (i.e. they equalled the target number) and used only the given numbers. In contrast to previous research (McMullen et al., 2016) answers were not coded as correct if brackets were omitted. Thus, $(0.75 + 0.25) \div 2 = \frac{1}{2}$ would be coded as correct, but $0.75 + 0.25 \div 2 = \frac{1}{2}$ would not. In McMullen et al.'s work with mid-elementary school students, flexible application of notational rules was understandable, since the younger students had not yet received instruction on the use of brackets and since arithmetic notation was not germane to the research. However, by the time students reach high school, they should be familiar with standard arithmetic notation. Therefore, it was decided that the most straightforward way of coding solutions was to evaluate the expressions precisely as written. If two correct expressions were identical, the solution was only counted once. However, mathematically identical solutions that used different notation (e.g. $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and $0.25 + 0.25 = \frac{1}{2}$) or varied the order of terms (e.g. $30 + 30 - 1 = 59$ and

$30 - 1 + 30 = 59$) were counted as distinct expressions. Each correct answer was given one point and sum scores were calculated for each item.

Reliability was good: Cronbach's α across all four items was 0.81; for only the whole number items Cronbach's α = 0.70 and for only the rational number items Cronbach's α = 0.76. Cronbach's alpha is intended for assessing consistency within a single-construct scale, but a high value does not necessarily imply that a scale *is* unidimensional (Taber, 2018). Whether the constructs of ANK and ARNK are distinct will be examined in Chapter 3. However, since all three statistics are fairly high, we can confidently consider this ASP task to be a reliable measure of adaptive number knowledge, whether we regard it as comprising two separate scales (ANK and ARNK) or not.

To facilitate further analysis, correct expressions were also coded according to whether they were (a) *complex* and (b) *cross-notational*. An expression was considered complex if it contained both additive (addition and/or subtraction) and multiplicative (multiplication and/or division) operations. An expression was considered cross-notational if it contained both common fractions and decimals. This second coding was only relevant for the rational number items. As these codings were not subjective, they were not checked for inter-rater reliability.

2.2.2 Missing Symbol Task

The Missing Symbol task consisted of arithmetic expressions in which participants had to fill in one or two missing numbers and/or operation symbols. This task was less open than the ASP task, in the sense that there was typically only one correct answer per item, but it was still expected to engage the creative and interconnected thinking processes that characterise adaptive number knowledge. Filling in missing numbers and (particularly) operations would require participants to think about the relationship between the numbers and operators that appeared in the expression. Some items had two blank spaces, which would require the participants to consider multiple possible combinations rapidly in their heads. However, it differed from the ASP task in the direction of thinking: to succeed in the ASP task, participants had to think "forwards" from the given numbers to a complete solution; in the Missing Symbol task, they needed to think "backwards" from an almost-complete expression back to the numbers and operators that might fit the blanks. Thus, in the ASP task participants could set

their own direction, potentially choosing to use strategies that they found easier, while in the Missing Symbol task they would have to conform to the approach dictated by the question.

The Missing Symbol task used in this test consisted of two sets of items. The first set consisted of eight items (arithmetic expressions) which used exclusively whole numbers. The target numbers in these sentences were 59 and 38. The second set consisted of eight items which used rational numbers as well as whole numbers. The target numbers in these sentences were $\frac{1}{2}$ and 3. All items used the same four target numbers as the Arithmetic Sentence Production Task described in Section 2.2.1, so that the results across both tasks would be as comparable as possible. An example item is shown in Figure 2.3.

Figure 2.3

Example Item from the Missing Symbol Task

The instructions read: "Fill in the missing numbers or operations as quickly as possible."

First, the participants were given one minute to complete a practice item containing two number sentences using small, familiar numbers ($6 \_\_ 7 = 42$ and $20 - \_\_ = 19$). After completing the practice item, they were invited to ask questions about the task. After this, they completed the test items. Participants were given three minutes for each set of eight number sentences. See Table 2.5 for a summary of test items.

Answers were coded as correct if the entire number sentence was mathematically correct (i.e. it equalled the target number). A correct answer was awarded 1 point and an incorrect answer was awarded 0 points. This meant that if there were two blank spaces in a number sentence, both needed to be filled in correctly to get the point. During coding, it became apparent that there were two possible correct answers for R4: $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$ and $2 \cdot \frac{1}{2} - \frac{1}{2} = \frac{1}{2}$ are both correct. Full credit was awarded for either answer. It was also noticed that R8 had no correct answer, so the results for R8 were excluded from subsequent analysis.

Table 2.5

Summary of Missing Symbol Task Test Items

Items labelled with a W are whole number items and items labelled with a R are rational number items.

Item	Number sentence	Number of blank spaces	Number or operation missing
W1*	$11 \cdot 6 \_\_ 7 = 59$	1	Op
W2*	$2 \_\_ 30 \_\_ 1 = 59$	2	Op, Op
W3	$11 \cdot 5 + \_\_ \cdot 2 = 59$	1	Num
W4	$12 \cdot \_\_ - 6 \_\_ 5 = 59$	2	Num, Op
W5	$15 \_\_ 3 - 7 = 38$	1	Op
W6	$3 \_\_ 6 + 4 \_\_ 5 = 38$	2	Op, Op
W7	$2 \cdot \_\_ + 2 \_\_ 9 = 38$	2	Num, Op
W8	$4 \_\_ 10 \_\_ 2 = 38$	2	Op, Op
R1	$0.75 \_\_ \frac{1}{4} = \frac{1}{2}$	1	Op
R2	$0.25 \_\_ \frac{1}{2} = \frac{1}{2}$	1	Op
R3	$\frac{1}{4} \div (\frac{3}{4} \_\_ \frac{1}{4}) = \frac{1}{2}$	1	Op
R4	$2 \_\_ \frac{1}{2} \_\_ \frac{1}{2} = \frac{1}{2}$	2	Op, Op
R5	$\frac{3}{2} \_\_ 1.5 = 3$	1	Op
R6	$0.75 \_\_ \frac{1}{4} = 3$	1	Op
R7	$\frac{3}{2} \div \_\_ + 2 = 3$	1	Num
R8**	$0.75 \_\_ \frac{3}{4} \cdot 2 = 3$	1	Op

*Items W1 and W2 were found to have very little variation and thus excluded from subsequent analyses.

**Note that there is no possible correct answer for R8 and thus it was excluded from all analyses.

Preliminary analysis showed that items W1 and W2 showed very little variation, with almost all participants answering them correctly, so they were also excluded from further analysis (see Section 5.1.2 for a full explanation). Sum scores were calculated for each set of items. Reliability was not as good as had been hoped: Cronbach's α across the six whole number items was 0.68 and for the seven rational number items Cronbach's α = 0.66. This figure is a little low. However, every item is dichotomous (i.e. the possible range of values is small),

which would tend to reduce inter-item correlations and therefore also Cronbach's α. Given that context, it seemed reasonable to consider 0.68 and 0.66 as reasonably high levels of internal consistency, rather than red flags. Across all 13 items, Cronbach's α = 0.76. It is normal for Cronbach's α to increase as more items are added to a scale (Taber, 2018), so it is not surprising that the statistic is higher when calculated across all test items.

2.2.3 Rapid Verification Task

The Rapid Verification task required participants to rapidly verify whether or not certain arithmetic expressions were equal to a given target number. This task was much less open than the ASP task and the Missing Symbol task, being very similar to routine classroom exercises in which students are asked to calculate the value of an arithmetic expression. However, it was suspected that when such tasks were done under high time pressure, they would encourage the activation of A(R)NK in order to rapidly verify whether an expression was or was not equal to the target number. For instance, if the target number were 59, a student might look at the expression $3 \cdot 22 - 6$ and rapidly reject it on the basis that the answer would be even, without needing to calculate the exact answer. Similarly, a student might look at the expression $180 \div 2 - 1$ and estimate that it would be near 90, therefore too large, and not calculate the exact answer before rejecting it. Utilising these shortcuts would make it much more likely that a participant would be able to complete all or most of the items in the given time, therefore high scorers would be likely to have the awareness of number properties, magnitude and operations which characterises A(R)NK.

The Rapid Verification task used in this test comprised four timed trials, each with a different target number (59, 38, ½ and ¾). The trials with target numbers 59 and 38 each contained 12 items (arithmetic expressions), which included exclusively whole numbers. The trials with target numbers ½ and ¾ each contained eight items and included a combination of rational and whole numbers.[5] In each trial, the participants were instructed to circle all the items that equalled the target number, and to work as quickly as they could. Participants were given one

[5] The rationale for having fewer items in the rational number trials was that rational number arithmetic tends to be more challenging for students and takes longer for them to complete, thus a smaller number of items was required to achieve the same time pressure.

minute to complete each trial. An example trial is shown in Figure 2.4. A summary of task items is given in Table 2.6. No practice items were given for this task.

Figure 2.4

Example Trial from the Rapid Verification Task

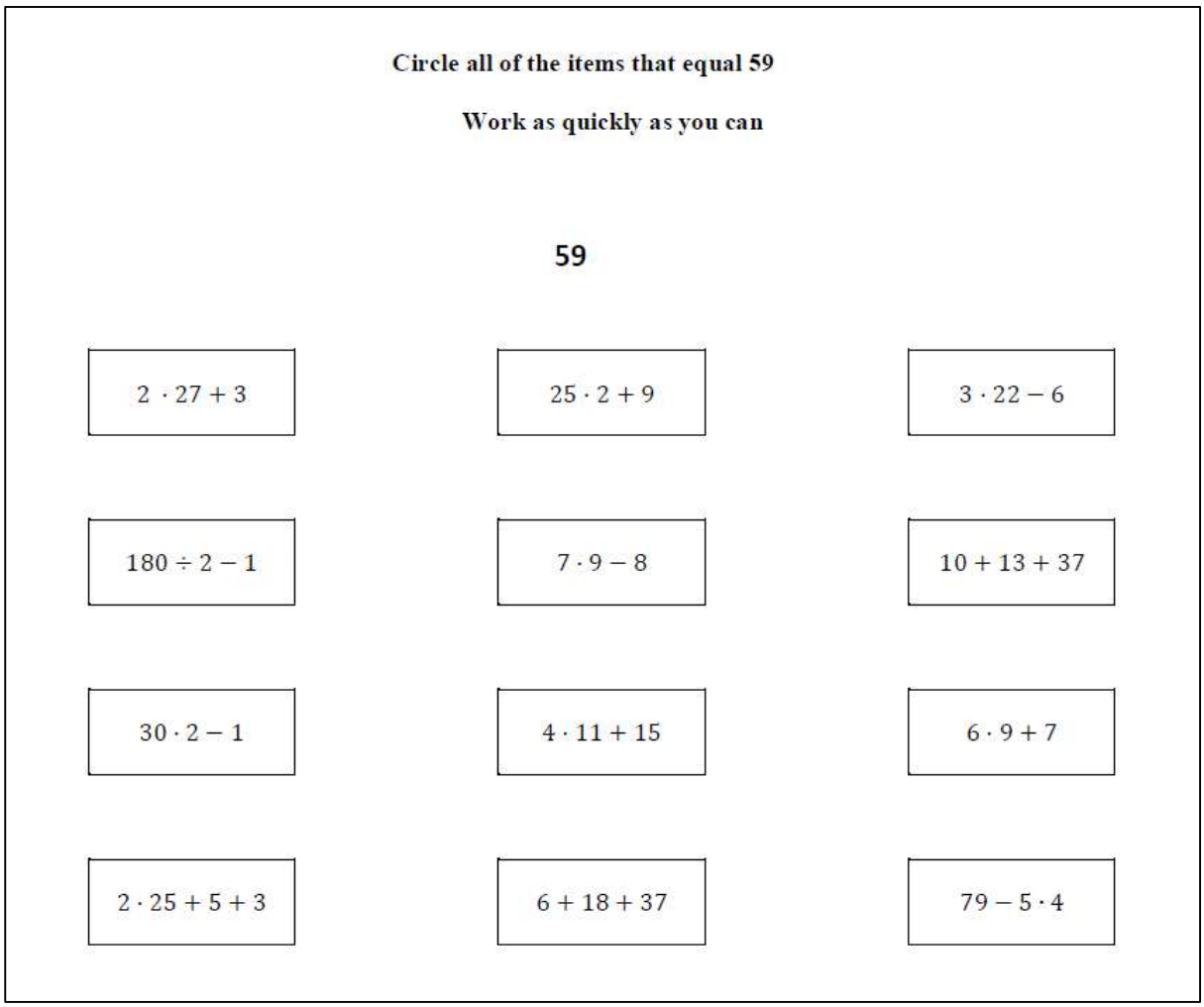


Table 2.6

Summary of Rapid Verification Task Items

Item	Expression	True/False	Item	Expression	True/False
59a	$2 \cdot 27 + 3$	False	38a	$5 \cdot 7 + 3$	True
59b	$180 \div 2 - 1$	False	38b	$4 \cdot 9 + 4 - 2$	True
59c	$30 \cdot 2 - 1$	True	38c	$2 \cdot 10 + 2 \cdot 9$	True
59d	$2 \cdot 25 + 5 + 3$	False	38d	$5 + 32 + 3$	False
59e	$25 \cdot 2 + 9$	True	38e	$14 \cdot 2 + 11$	False
59f	$7 \cdot 9 - 8$	False	38f	$70 \div 2 - 3$	False
59g	$4 \cdot 11 + 15$	True	38g	$3 \cdot 10 - 4 \div 2$	False
59h	$6 + 18 + 37$	False	38h	$9 \div 3 \cdot 12$	False
59i	$3 \cdot 22 - 6$	False	38i	$3 \cdot 20 - 2$	False
59j	$10 + 13 + 37$	False	38j	$80 \div 2 - 2$	True
59k	$6 \cdot 9 + 7$	False	38k	$3 \cdot 6 + 4 \cdot 5$	True
59l	$79 - 5 \cdot 4$	True	38l	$16 \cdot 2 + 6$	True
½ a	$0.5 \div \frac{1}{4}$	False	¾ a	$\frac{1}{2} + \frac{1}{4}$	True
½ b	$0.25 + 0.25$	True	¾ b	$1 - 0.25$	True
½ c	$\frac{1}{4} \div 0.5$	True	¾ c	$\frac{1}{3} + 0.25$	False
½ d	$(0.75 - 0.5) \cdot 2$	True	¾ d	$0.25 \div \frac{1}{3}$	True
½ e	$\frac{3}{4} - \frac{1}{4}$	True	¾ e	$0.5 - 0.25$	False
½ f	$0.75 \cdot 0.25$	False	¾ f	$\frac{1}{4} + \frac{1}{4}$	False
½ g	$075 - \frac{1}{4}$	True	¾ g	$\frac{1}{4} \div \frac{1}{3}$	True
½ h	$\left(\frac{3}{4} + \frac{1}{4}\right) \div \frac{1}{2}$	False	¾ h	$\frac{1}{4} + 0.25 + \frac{1}{4}$	True

In the first trial (target 59), four out of 12 options were equal to the target number. In the second trial (target 38), six out of 12 options were equal to the target number. In both the third and fourth trials (target ½ and target ¾), five out of eight options were correct. This meant that in total there were 10 correct answers out of 24 whole number items and 10 correct answers out of 16 rational number items.

Initially, answers were coded as correct if the participant had correctly circled an item that equalled the target number (henceforth called a *True item*) or had correctly left blank an item that did not equal the target number (henceforth called a *False item*). Answers were coded as incorrect if the participant had incorrectly circled a False item or incorrectly left blank a True item. A correct answer was awarded 1 point and an incorrect answer was awarded 0 points.

The reliability for each timed trial was relatively low, with Cronbach's α ranging from 0.37 to 0.60 (see Table 2.7). One contributing factor to the low reliability was probably the binary nature of each item, which reduced the possible variation and therefore also the inter-item correlations on which Cronbach's α was based. It was thus decided to calculate a sum-score for each trial (possible range: 0-12 for whole number trials, 0-8 for rational number trials) and use these to calculate reliability statistics for the whole number trials, the rational number trials and the Rapid Verification task overall. The logic here was that the sum-scores would have greater variation, thus making potential correlations easier to detect. The new reliability statistics were noticeably higher (0.60 for whole trials, 0.71 for rational trials, 0.76 overall), indicating a reasonable amount of internal consistency.

Table 2.7

Reliability statistics for Rapid Verification Task

Scale	**Component Items**	**Cronbach's α**
Target 59	12 individual items (arithmetic expressions) with target 59	0.37
Target 38	12 individual items (arithmetic expressions) with target 38	0.47
Target ½	8 individual items (arithmetic expressions) with target ½	0.38
Target ¾	8 individual items (arithmetic expressions) with target ¾	0.60
Whole	2 sum-scores from the whole number trials	0.60
Rational	2 sum-scores from the rational number trials	0.71
All	4 sum-scores from all trials	0.76

However, binary items could not entirely explain the low reliability levels, since the items in the Missing Symbol task were also binary, and that task had higher reliability levels than the Rapid Verification task. This suggested that there was a further source of inconsistency within the Rapid Verification task responses. A possible explanation for that inconsistency might be

that respondents seemed to answer True items in a different way to False items, as outlined in the paragraphs that follow.

During preliminary analysis, it was noted that a simple correct/incorrect coding seemed to be insufficient. The correct rates for the False items were noticeably higher than the correct rates for the True items. It therefore seemed possible that participants had a bias against circling answers. The framework of signal detection theory (SDT) was used to generate insight into participants' decision-making processes. SDT is a framework which helps researchers to understand not only the accuracy of answers in a testing process, but also the decision-making strategies behind the answers (Abdi, 2010; MacMillan, 2002). It has been used previously in educational research, for instance by Geary et al. (2009). It is a useful framework because in a multiple-choice (or binary choice) scenario it allows one to distinguish a student who gets many correct answers by circling only the correct items from a student who gets many correct answers by circling everything – and similarly allows one to distinguish students who seem to actively reject incorrect answers from those who do not answer anything at all. In Geary et al.'s (2009, p. 271) words, "[t]he analysis enables separation of children's sensitivity to [the variable of interest] from their tendency to respond (i.e., circle) to test items".

In accordance with SDT, answers were coded as hits, misses, false alarms (FAs) or correct rejections (CRs) (Abdi, 2010; Geary et al., 2009; MacMillan, 2002). A True item that was circled was coded as a hit; a True item that was not circled was coded as a miss. A False item that was circled was coded as a false alarm; a False item that was not circled was coded as a correct rejection. Sum scores were calculated for the number of hits, misses, FAs and CRs for each trial separately, for the whole number trials and the rational number trials separately, and for all trials jointly. In addition, the total number of correct answers of any kind (hits or CRs) was recorded for each separate trial, whole number trials and rational number trials, and all trials jointly. These values will be utilised in Chapter 5, when the possibility of a systematic response bias in the Rapid Verification task is considered.

2.3 Statistical Analysis

The data were explored using a range of descriptive and inferential statistical techniques, which are described in the chapters that follow. All analyses were conducted in IBM SPSS Statistics

Version 27, except the chi-square analyses described in Section 5.2, which were conducted by hand in Microsoft Excel 2019.

3. The Relationship Between ANK and ARNK

This chapter addresses the first research question, namely: *Are adaptive whole number knowledge and adaptive rational number knowledge empirically distinct constructs?* In order to answer this question, only the data from the arithmetic sentence production task were considered. This was because the ASP task is the single established measure that exists for ANK[6] and ARNK. The data were used to investigate whether the whole number items and rational number items seemed to be measuring a single construct. If the answer was yes, that would have suggested that adaptive whole number knowledge and adaptive rational number knowledge are not distinct. If, on the other hand, the answer was no, that would have provided support for the hypothesis that adaptive whole number knowledge and adaptive rational number knowledge are distinct constructs.

The data were analysed in two ways. Firstly, inter-item correlations and internal consistency were calculated and examined. Secondly, an exploratory factor analysis was conducted to investigate whether latent factors emerged which could be identified with ANK, ARNK, or both. Four variables were considered in these analyses: the total number of correct responses on each of the two whole number items (target numbers 59 and 38), and the total number of correct responses on each of the two rational number items (target numbers ½ and ¾).

3.1 Inter-Item Correlations and Internal Consistency

Pearson correlation coefficients were calculated for each pair of variables, excluding cases listwise in the event of missing data. As can be seen in Table 3.1, the correlation coefficients ranged from $r = 0.45$ (between target number 38 and target number ½) to $r = 0.61$ (between target number ½ and target number 3), $p < 0.001$ for all calculated correlation coefficients. These are moderately strong positive correlations, so it is plausible that the two variables in each pair are measuring aspects of the same latent factor(s) – potentially some factor of general adaptive number knowledge. Another observation in support of this interpretation is that the higher correlations did not occur exclusively between items of the same type (whole-whole and

[6] The abbreviation "ANK" is used in this chapter to refer specifically to adaptive *whole* number knowledge. The full form "adaptive number knowledge" is used to refer to adaptive number knowledge in general, unless specified otherwise.

rational-rational): the second-highest correlation, $r = 0.58$, was found between a whole number item (target number 59) and a rational number item (target number 3).

Table 3.1

Pearson Correlations Between Items in the Arithmetic Sentence Production Task

	Total Correct: Target 59	**Total Correct: Target 38**	**Total Correct: Target 1/2**	**Total Correct: Target 3**
Total correct: Target 59	--			
Total correct: Target 38	0.55***	--		
Total correct: Target ½	0.53***	0.45***	--	
Total correct: Target 3	0.58***	0.54***	0.61***	--
*** Correlation is significant at the 0.001 level (2-tailed).				
Listwise N=405				

On the other hand, the correlations were not so high as to suggest that any of the variables were measuring identical things. This is important for two reasons. Firstly, because it leaves open the possibility that ANK and ARNK are distinct constructs. It may be that the correlation between the whole number items and the rational number items can be explained by some other factor – for instance, general intelligence (although this is unlikely, since McMullen et al., 2019 & 2022, did not find a relationship between domain-general predictors and ANK or ARNK), procedural fluency or general mathematics achievement. Secondly, because extreme multicollinearity is problematic for many types of exploratory factor analysis: extreme multicollinearity makes it difficult to establish the unique contributions that highly correlated variables make to factors (Field, 2009). Since the Pearson correlation coefficients indicate there is *not* extreme multicollinearity between any of the variables considered, this finding fulfils one of the prerequisites for conducting a factor analysis. (Other prerequisites will be discussed in Section 3.2.)

Cronbach's alpha for these four variables was 0.81, which suggested a fairly high degree of internal consistency. A high value of alpha does not guarantee that a set of items is unidimensional, but it does indicate "that every item … [is] measuring something similar to *some* of the other items" (Taber, 2018, p. 1286). This suggests that there is at least some commonality between the whole number and rational number tests of adaptive number knowledge, which in turn might indicate that the ANK and ARNK are a single construct.

Both the correlations and Cronbach's alpha suggested that there may be a single latent factor which could explain the variance in both ANK and ARNK scores. Therefore, it was decided to conduct an exploratory factor analysis to investigate the matter further.

3.2 Exploratory Factor Analysis

3.2.1 Preliminary checks on distributional properties of the data

Exploratory factor analysis (EFA) assumes that the latent constructs exert a linear influence on measured variables (Watkins, 2018). The linear correlations between these variables – which are used to compute the results of a factor analysis – may be influenced by the distributional properties of the data (specifically variability, normality and linearity), and so it is important to investigate the relevant distributional properties before proceeding with an EFA.

Variability

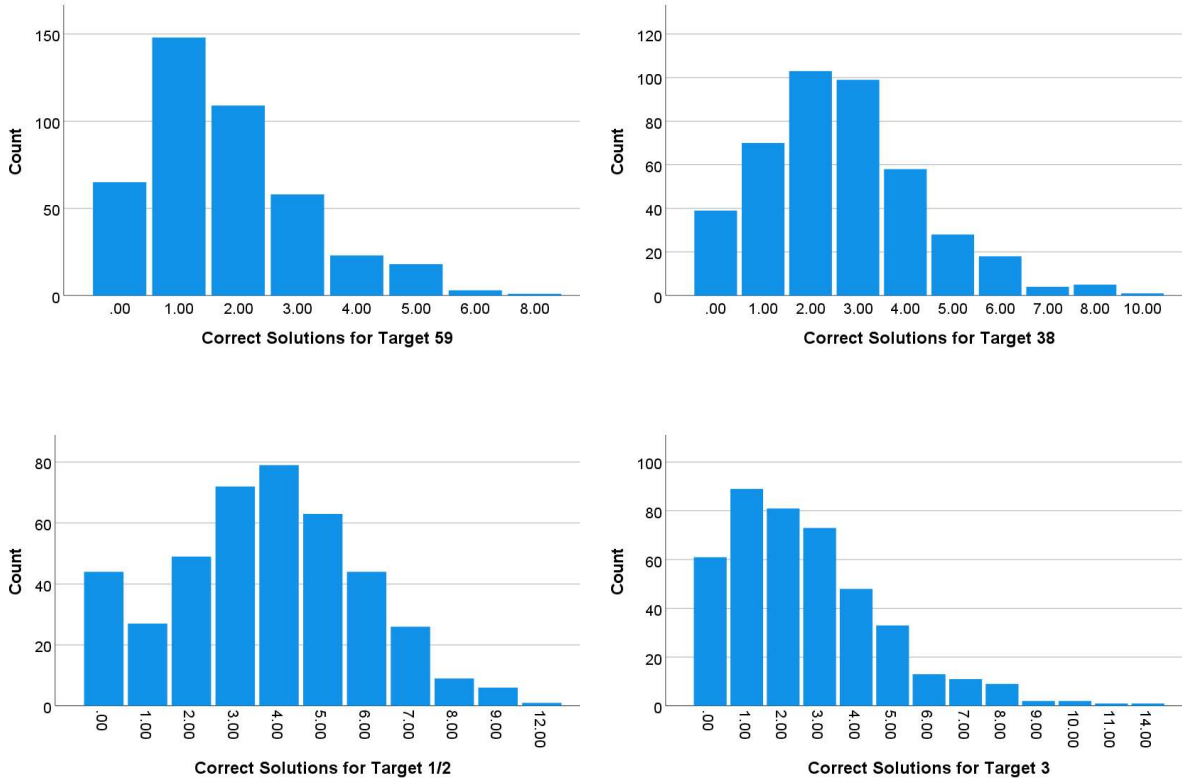
Watkins (2018) advises checking that the range of the variables of interest has not been attenuated by the sample selection process. A sample which is much more restricted than the population will also have restricted variance. This may attenuate the correlation coefficients, obscuring a relation which does exist in the broader population.

The present sample contained a large number of students in Honors courses (i.e. mathematics courses which are more demanding than standard courses) and Advanced Placement courses (i.e. courses which teach college-level mathematics, and which are thus more advanced than typical school-level mathematics courses). In total, 368 out of 447 participants (82%) were enrolled for Honors or AP courses. In addition, 70.2% of participants received either an A or a B for their latest mathematics grade, which are the two highest grades. This might suggest that the sample has a level of mathematical ability which is higher than the population mathematical ability, although this is impossible to confirm without data on the course enrolment statistics and mathematics grades of other students in the population. ANK is considered to be a characteristic of high-level mathematics performance (McMullen et al., 2019), so it is possible that in a sample that is (potentially) highly mathematically talented, A(R)NK scores may be clustered at the high end of the scale.

In order to investigate this possibility, histograms of the score distribution for each variable of interest were generated (see Figure 3.1). From the histograms it appears that there is no clustering of scores at the upper end of the range; in fact, the scores seem to follow roughly normal distributions in which neither the lowest nor highest scores are the most common. There is a definite skew to the right on most items, but the clustering of scores at the lower end of the range is not so severe as to expect it to render a factor analysis ineffective.

Figure 3.1

Histograms Showing Score Distributions for ASP Task Items



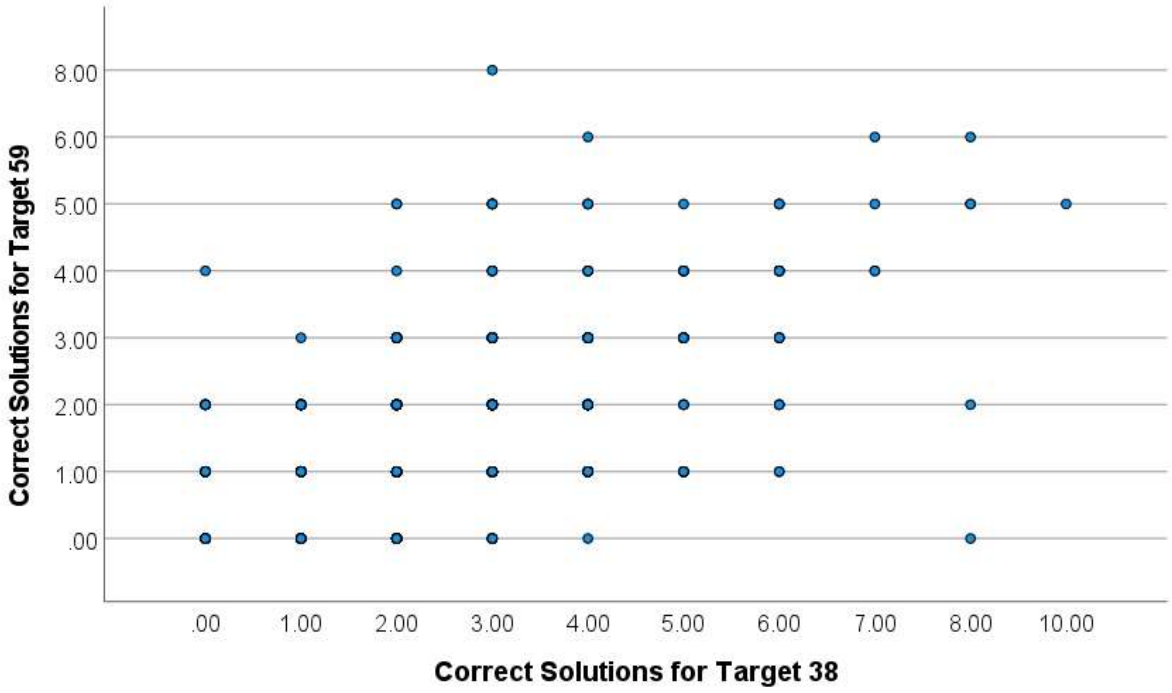
Linearity

Watkins (2018) states that since Pearson correlation coefficients measure a linear relationship, they will be reduced if the relationship between the variables is not actually linear. Watkins further states that linearity can be subjectively judged by examining scatter plots of the variables. Simple scatter plots were initially generated, but because the number of correct

solutions is a discrete variable, the resulting scatter plots looked like simple grid arrangements of dots, with no way to tell how many dots were stacked on top of each other in each location. This makes it difficult to tell if any given dot is an outlier. An example of such a scatter plot is given in Figure 3.2. In order to make the scatter plots easier to interpret, they were regenerated using the JITTER command. This command "[a]dds a small amount of random noise to all scale axis dimensions … allowing separation of coincident points" (IBM Corporation, 2021d). Jittered scatter plots comparing all four of the variables to each other are given in Figure 3.3.

Figure 3.2

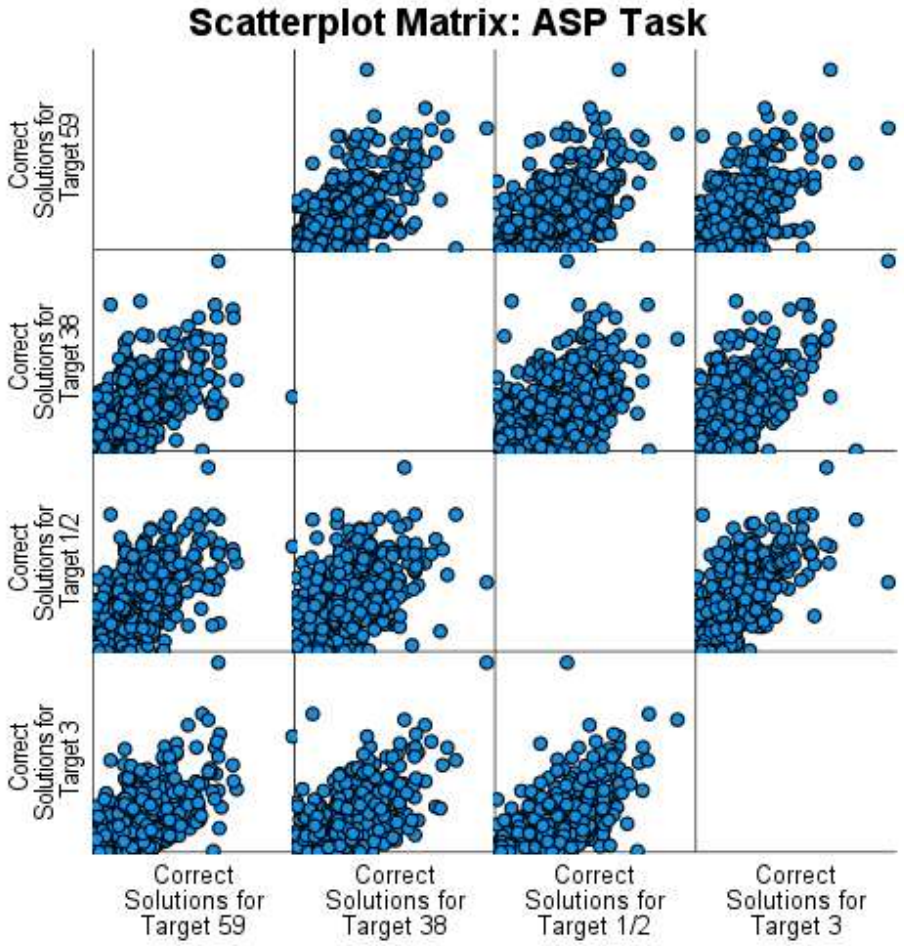
Simple Scatter Plot Showing the Relationship Between Whole Number ASP Task Items



From these scatter plots, it is clearly apparent that there is a positive linear relationship between each pair of variables. The relationship is not extremely strong, which can be seen from the fact that the dots form diagonal "clouds" rather than a very obviously line-like pattern, but it is nonetheless apparent at a glance. This is consistent with the Pearson correlation coefficients calculated in Section 3.1, which showed moderate positive linear correlations between all the measured variables.

Figure 3.3

Jittered Scatter Plots Showing the Relationship Between All ASP Task Items



Normality

Watkins (2018) states that skewness and kurtosis have a particularly strong effect on the Pearson correlation coefficient,[7] and therefore on EFA results. Curran et al. (1996) found serious problems in a series of simulations when univariate skewness $\geq$ 2.0 and univariate kurtosis $\geq$ 7.0. Curran et al. therefore suggest that skewness and kurtosis values approaching these numbers be regarded as suspect.

As discussed above, the histograms (Figure 3.1) suggested that the relevant variables were distributed approximately normally, but with a noticeable skew to the right. The responses for

[7] It should be noted that if the data is not linear or is not normal, an EFA does not necessarily need to be abandoned; other types of correlation coefficients can be considered as an alternative (Watkins, 2018).

target number 38 and target number ½ appeared more normal than the responses for target number 59 and target number 3. In the latter cases, a larger number of responses were clustered at the lower end of the range, which made the normal curve look somewhat truncated on the left. However, none of the histograms appeared to be decisively non-normal.

In order to examine the question of normality with more precision, descriptive statistics were examined. In Table 3.2, it can be clearly seen from the location of the mean values and the size of the standard deviation (SD), relative to the minimum and maximum statistics, that all the distributions of all four variables are skewed to the right. In all cases, the minimum value lies between one and two standard deviations below the mean, while the maximum value is multiple standard deviations above the mean.[8] This is consistent with the interpretation of the histograms.

Table 3.2

Descriptive Statistics for ASP Task Items

	N	Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
Total correct: Target 59	425	0	8	1.76	1.36	0.98	1.13
Total correct: Target 38	425	0	10	2.66	1.72	0.73	0.84
Total correct: Target 1/2	420	0	12	3.70	2.19	0.17	- 0.16
Total correct: Target 3	424	0	14	2.63	2.16	1.22	2.29
Valid N (listwise)	**405**						

However, just as the histograms suggested, the non-normal tendencies are not extreme. Both the skewness and kurtosis statistics are well within the bounds recommended by Curran et al. (1996). The largest figure for skewness is 1.22, which is well under the "red flag" value of 2.0, and the largest figure for kurtosis is 2.29, which is similarly well under 7.0.[9] Both of the largest figures are found in the distribution of the final ASP item (target number 3).

[8] For target number 59, the maximum value is 4.6 SD above the mean. For target number 38, the maximum value is 4.3 SD above the mean. For target number ½, the maximum value is 3.8 SD above the mean. For target number 3, the maximum value is 5.3 SD above the mean.

[9] Mishra et al. (2019) suggest a stricter criterion for absolute kurtosis, preferring a value under 4. The data under consideration also meets this more stringent standard.

A final test of normality was conducted by means of the Kolmogorov-Smirnov test. This normality test was selected because Mishra et al. (2019) state that the Kolmogorov-Smirnov test is appropriate for samples with n $\geq$ 50, while the Shapiro-Wilk test is appropriate when the sample size is smaller. The Kolmogorov-Smirnov test indicated that all four scoring distributions deviated significantly from normality: for target number 59, $D(425) = 0.21$, $p = 0.000$; for target number 38, $D(425) = 0.15$, $p = 0.000$; for target number ½, $D(420) = 0.10$, $p = 0.000$; and for target number 3, $D(424) = 0.16$, $p = 0.000$.

It is not surprising that the Kolmogorov-Smirnov test detected significant deviation from normality, because the sample size is large. Statistical tests of normality can be overly sensitive to large sample sizes (Field, 2018; Mishra et al., 2019). Exactly what counts as a large sample size is hard to determine, as it depends on the exact nature of the sample: whether it is light-tailed or heavy-tailed, and whether it has extreme skew or kurtosis (Field, 2018). However, Field (2018) suggests that n > 160 would be considered large for a heavy-tailed distribution, and Mishra et al. (2019) regard n > 300 as large. Thus, it seems reasonable to conclude that the present sample sizes of n $\geq$ 420 are sufficiently large not to be overly concerned by the results of the Kolmogorov-Smirnov tests, given that a visual inspection of the histograms as well as the skewness and kurtosis statistics indicate that the distribution is sufficiently normal to proceed with an EFA.

3.2.2 Preliminary check on sample size

An important question to consider is whether the sample size is sufficiently large to carry out an EFA. Watkins (2018) does not give concrete guidelines what counts as sufficiently large, because the answer is affected by several factors, including the number of measured variables, the ratio of variables to factors, and the level of communality – not all of which can be known in advance, given that factor analyses are often exploratory (Mundfrom et al., 2005). However, Mundfrom et al.'s (2005) simulation study provides more specific guidance. All else being equal, the smaller the number of input variables, the larger a sample is needed for good results. More specifically, a high number of variables per factor markedly decreases the required sample size. Therefore, if one has an idea of how many factors are expected, it is possible to approximate a minimum sample size. In the present case, the most likely outcomes of an EFA would be a one-factor solution (a single ANK construct across both whole and rational

numbers) or a two-factor solution (distinct ANK and ARNK constructs). Since there are only four indicator variables, this suggests a variable-to-factor ratio of 4:1 in the former case and 2:1 in the latter.

In the most demanding scenario (low level of communality,[10] excellent agreement between sample and population solutions), a sample size of $n = 95$ was found to be sufficient for a 4:1 ratio with a single-factor solution (Mundfrom et al., 2005). Therefore, the present sample size of $n = 405$ would be more than sufficient if there is one latent factor. However, if there are two factors in the present data, meaning only two variables per factor, a larger sample size would be needed. Mundfrom et al.'s analysis only investigated scenarios starting from a minimum of three variables per factor.[11] In the most demanding scenario (low level of communality, excellent agreement between sample and population solutions), a minimum sample size of $n = 900$ was determined to be necessary for a 3:1 ratio with a two-factor solution. This is clearly larger than the sample in the present study, and in fact even more than 900 participants would be required with a lower variable-to-factor ratio of 2:1. However, if the least demanding scenario is considered (high level of communality, merely good agreement between sample and population solutions), then a sample size of $n = 90$ would suffice for a 3:1 ratio with a two-factor solution. Even adjusting the sample size up somewhat to allow for a 2:1 ratio, the present sample would be comfortably large enough in this scenario. In a middle-ground scenario (wide level of communality, good agreement between sample and population solutions), a sample size of $n = 160$ was found to be sufficient for a 3:1 ratio with a two-factor solution. This also suggests that the present $n = 405$ sample would be sufficient in the middle-ground scenario.

Therefore, it is likely that the present data can support an EFA. This is certainly true if the EFA reveals a one-factor solution. However, if a two-factor solution is uncovered, the communalities should be examined carefully. A two-factor solution with low communalities should be regarded with caution, since the sample size may not be large enough to produce a trustworthy result in those circumstances.

---

[10] Mundfrom et al. tested models with three patterns of communality. "High" meant that communalities were between 0.6 and 0.8, "wide" meant that communalities were between 0.2 and 0.8, and "low" meant that communalities were between 0.2 and 0.4.

[11] Watkins (2018) notes that ideally, at least three variables per factor are required to identify the factor. Statistical software can compute a solution with fewer than three factors per variable, but the solution is likely to be imprecise.

3.2.3 Preliminary checks on other aspects of the data

Watkins (2018) also identifies certain other guidelines for determining whether a data set is suitable for an exploratory factor analysis. These further guidelines will be considered in this section.

First, the data should be measured on a *ratio or interval scale of measurement*, because that is assumed by Pearson correlations. The present data uses an interval scale, so this is not problematic.

Second, the quantity and nature of any *missing values* should be considered. No more than 6% of the data was missing from any of the ASP items (given that the total sample size was $n = 447$ and the smallest number of participants for whom valid responses were recorded was 420, for target number ½). In total, 42 participants (9.4%) were missing data for at least one of the ASP items. In coding the ASP task, zero was regarded as a valid number of solutions, so a missing value did not indicate that a student produced no correct solutions. Rather, it meant that none of their solutions were legible. This was usually the result of a poor-quality scan: the hard-copy test scripts were scanned in the United States and the scans were emailed to Finland. By the time illegible scans were noticed, it was not practical to obtain better scans of those scripts. The illegible scans seemed to be randomly distributed throughout the sample. Sometimes only one or two pages of a script were illegible, while the rest of the script could be made out. Therefore it is likely that there is no systematic pattern to the missing data. For this reason, excluding the missing data seems unlikely to cause trouble. Following Field's (2009) recommendation, cases with missing data were excluded listwise for the EFA which follows, which still left a large sample of $n = 405$.

Third, *outliers* should be checked to ensure there are no out-of-range values or missing-value codes masquerading as real data points. This was checked and no such outliers were found. In addition, valid outlying values might come from a population that differs from the intended population for the study (Watkins, 2018). Since the present distributions are skewed to the right, all values greater than 2 SDs away from the mean lie on the right of the mean. However, this is not a cause for concern in a study of A(R)NK. Previous research on this subject

(McMullen et al., 2017, 2019, 2020) has found considerable individual differences in scores on the ASP task, with the smallest cluster of students having very high scores. Therefore, outliers at the upper end of the distribution are to be expected, and are unlikely to be representatives of some other population. Thus, all cases were retained.

Fourth, if possible, variables used in an EFA should have *reliability* of at least 0.70. Watkins (2018) states that lower reliability would leave little variance to be shared with other variables. Watkins does not go into detail regarding why this is the case, but presumably it is because (as outlined in Section 3.1) a high value of Cronbach's alpha – a common measure of reliability – does not guarantee that a set of items is unidimensional, but it does indicate that each item in the set is correlated with at least some of the other items in the set (Taber, 2018). Thus a high Cronbach's alpha score would indicate that there are one *or more* cluster(s) of items which return similar answers – and if there are, that in turn suggests that there are common underlying factors which might be found by an EFA. In the present data, this criterion was met: as reported in Section 3.1, Cronbach's alpha for the subscale consisting of the sum scores for all four ASP task items was high, at 0.81.

Finally, it is wise to make a *general assessment* of whether the variables are correlated to the extent that makes it likely there is some systematic shared variance (i.e., latent factors). Watkins (2018) explains that this can be done subjectively and objectively. Subjectively, by examining the inter-item correlation matrix, as was done in Section 3.1. Objectively, with Bartlett's test of sphericity, which tests the hypothesis that the correlation matrix was generated by random data. If Bartlett's test generates a statistically significant χ^2 value, the correlation matrix is unlikely to be random and therefore an EFA is justified. Since Bartlett's test is very sensitive with large sample sizes, Watkins recommends that it be supplemented in these cases by the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. The KMO measure "reflects the extent to which correlations are a function of the variance shared across all variables rather than the variance shared by particular pairs of variables" (Watkins, 2018, p. 226). Watkins reports a general consensus that the KMO should ideally be greater than or equal to 0.7. Values less than 0.5 are considered unacceptable and suggest that a factor analysis is unlikely to produce useful results (IBM Corporation, 2021b; Watkins, 2018).

Bartlett's test of sphericity indicated that the correlation matrix was significantly different to the identity matrix, $\chi^2(6, n = 405) = 576.24, p < 0.001$. The KMO statistic was 0.80, which

is well above the recommended minimum level of 0.5, and also exceeds the desired benchmark of 0.7. All KMO values for individual items were greater than 0.77, which is above the acceptable limit of 0.5 (Field, 2009). Taken together, these results provide further reason to believe that the data are well-suited to factor analysis.

3.2.4 Exploratory Factor Analysis or Principal Components Analysis?

While Principal Components Analysis (PCA) is often informally included under the umbrella term "exploratory factor analysis", the two types of analysis are in fact theoretically and computationally distinct (Watkins, 2018). PCA aims to reduce the original variables into linear combinations called *components*, which jointly explain as much variation as possible in the original variables, but which are not latent constructs – instead, components are deemed to be influenced *by* the measured variables (Watkins, 2018). A primary aim of PCA is data reduction, and the results should not be extrapolated beyond the sample in question (Field, 2018; Raykov & Marcoulides, 2011). By contrast, EFA was developed in order to identify latent *factors* and thus generate hypotheses for future research (Field, 2018; Raykov & Marcoulides, 2011). Notably, EFA computations separate out the variance that is unique to each variable from the variance that is shared with other variables, something that PCA does not do (Field, 2009; Watkins, 2018).

However, in practical application, the differences between EFA and PCA are minimal; they tend to result in the same overall factor/component structure, with minor differences in factor/component loadings (Field, 2018; Guadagnoli & Velicer, 1988; Velicer et al., 1982). This is recognised explicitly by Field (2009), who opts to focus on PCA rather than EFA in his theoretical discussion and worked example, on the grounds that it is psychometrically sound and conceptually less complex than EFA. (It should be noted, however, that in the 2018 edition of the same book, Field focuses the worked example and considerably more of the discussion on EFA.) It is also recognised implicitly by Raykov and Marcoulides (2011), who present a worked example of a PCA in SPSS in order to explore whether a latent factor can explain the response patterns in a battery of intelligence tests – even though identifying a latent factor is theoretically something which requires an EFA.

The choice of whether to utilise a PCA or an EFA was therefore deemed unlikely to affect the outcome of the analysis. However, for the sake of comparison, multiple analyses were conducted. Section 3.2.5 reports on the results of one PCA and two EFAs with different factoring methods: principal axis factoring, which does not assume normality and is sensitive to weaker factor-variable relations (Watkins, 2018), and maximum likelihood estimation, which does assume normality, but which allows one to generalise from the sample to the general population (Field, 2018; Watkins, 2018). These are the two most common factoring methods (Watkins, 2018).

3.2.5 Results of the PCA and Exploratory Factor Analyses

Principal Components Analysis

A principal components analysis was conducted on the four items, using an oblique rotation (promax). An orthogonal rotation would only be appropriate if the factors were expected to be independent (Field, 2009). However, it seems likely that if distinct factors of ANK and ARNK were detected, they would be correlated with each other, since they would require many of the same skills and the same understanding of whole number magnitude and calculation skills. Therefore, an oblique rotation was selected. In the event that independent factors are found, the promax rotation would produce orthogonal results (Watkins, 2018), so it was judged that an appropriate rotation would be obtained regardless of the nature of the latent factors.

The communalities were relatively high according to the guidelines used by Mundfrom et al. (2005), ranging from 0.60 to 0.72. This is a positive finding, since Raykov and Marcoulides (2011) state that high communalities indicate a well-fitting solution.

An initial analysis was run to obtain eigenvalues for each component in the data. Only one component had an eigenvalue that exceeded Kaiser's criterion of 1, and that component explained 65.82% of the observed variance in the indicator variables (see Table 3.3). The scree plot also clearly confirmed the existence of a single major component. As only one component was identified, the solution was not rotated.

Table 3.3

Initial Solution of Principal Components Analysis

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.63	65.82	65.82	2.63	65.82	65.82
2	0.56	14.07	79.88			
3	0.43	10.82	90.71			
4	0.37	9.29	100.00			

Table 3.4 shows the factor loadings on this single component, as well as factor loadings that were calculated using Principal Axis Factoring and Maximum Likelihood extraction (see below). All four variables loaded strongly onto the identified component, with loadings ranging from 0.78 to 0.85.

Table 3.4

Component/Factor Matrices for Three Different Extraction Methods

Indicator	Loading on Component 1 (PCA)*	Loading on Factor 1 (PAF)*	Loading on Factor 1 (ML)*
Total correct: Target 3	0.85	0.81	0.81
Total correct: Target 59	0.82	0.75	0.74
Total correct: Target ½	0.80	0.72	0.73
Total correct: Target 38	0.78	0.68	0.68

*PCA = Principal Components Analysis, PAF = Principal Axis Factoring, ML = Maximum Likelihood

Exploratory Factor Analyses

For the sake of comparison, two EFAs were conducted on the four items, also using the promax rotation. The first EFA used principal axis factoring and the second used maximum likelihood estimation. In both EFAs, the communalities after extraction were found to be lower than in the PCA described above, ranging from 0.46–0.65 in the case of principal axis factoring, and 0.46–0.65 in the case of maximum likelihood estimation. However, this is not surprising since factor analysis separates out unique variance from common variance (i.e. from communalities), while PCA does not. It is therefore to be expected that the communalities would be lower in an EFA of the same data.

Both the principal axis factoring and maximum likelihood EFAs resulted in identical initial solutions to the PCA reported above, with a single major factor explaining 65.82% of the total variance before extraction. After extraction, the major factor explained 54.65% of the variance (the percentage of variance explained was the same for both extraction methods). The scree plot confirmed a single-factor interpretation. Since only one factor was identified, the solution was not rotated.

Table 3.4 shows the factor loadings on this single factor for both factor analyses. Although the factor loadings are slightly lower than in the PCA, they are still fairly high, with loadings ranging from 0.68 to 0.81.

After comparing the outcome of the PCA to the outcomes of the EFA, it can be concluded that the choice of procedure does not affect the result.

3.3 Considering the Effects of Age, School Grade and Mathematics Module

The author initially intended to run similar EFAs on subsets of the data to assess whether the results were consistent across school grades, ages and mathematics modules. It is plausible, for instance, that lower-attaining students (who would be more likely to be enrolled in lower-level mathematics modules) would be more likely to struggle with fraction arithmetic and thus would generate fewer correct answers on the ASP task, even if they have developed a reasonable degree of competence with whole number arithmetic by the time they reach high school. This might mean that ANK and ARNK would be distinguishable in lower-achieving samples, but not in higher-achieving samples. It is similarly plausible that ANK and ARNK might diverge in younger students, who have learned about fractions more recently and may not have fully integrated rational and whole number knowledge, while converging in older students, who have had additional exposure to rational number applications in subjects like algebra and geometry, which could support the integration of whole and rational number knowledge.

However, the nature and size of the sample made these analyses impossible. As noted in Section 3.2.2, factor analysis requires a large number of cases to generate reliable results, and this is particularly true when the number of indicator variables is small. The present sample is

partitioned into many subgroups, which means that very few subgroups are large enough to perform a factor analysis with.

In the case of mathematics module, the only two subgroups which are large enough to perhaps consider an EFA are the students enrolled in Algebra 2 Honors ($n = 130$) and in Geometry Honors ($n = 167$). These numbers may be too small in the event of a two-factor solution (see Section 3.2.2), but more importantly, these are not useful mathematics modules to compare. It is possible to enrol in Algebra 2 Honors and Geometry Honors concurrently (see Table 2.2), and so it is entirely possible that some children in the sample are studying both modules.[12] In addition, Honors modules are intended for higher-achieving students, so there is unlikely to be a large difference in mathematical skill between the two groups. Thus, comparing these subgroups would be unlikely to reveal any differences.

In the case of school grade (which largely corresponds with age), the sample consists mainly of ninth- and tenth-grade students (or 14- and 15-year-olds). There are 236 ninth-graders and 114 tenth-graders, or alternatively 119 fourteen-year-olds and 169 fifteen-year-olds. The subsamples of older students are much smaller. Again, there are two problems with comparing these groups using an EFA. The first is size. If a two-factor solution is a possibility, these sample sizes are simply not big enough for the number of indicator variables. The only group that is potentially large enough is the ninth-grade subsample, but there would be no other group to compare it to. Furthermore, the results for ninth-graders are likely to mirror the whole-group results, since they make up such a large proportion of the total group. The second problem is that these larger groups are similar in age and grade, and therefore a comparison is unlikely to reveal major differences. A comparison between ninth-graders and twelfth-graders, for instance, would be more likely to be informative.

For these reasons, it was decided that factor analyses of subgroups should not be conducted. However, this is something that could be considered in future studies with larger samples.

[12] It should be noted that no participant in the study wrote the test twice. However, not all mathematics classes were tested, so a student who wrote the test in their Algebra 2 Honors classroom (and was recorded as an Algebra 2 Honors student) may also have a Geometry 2 Honors class on their timetable.

3.4 Discussion and Conclusions

The results above provide good evidence to support the hypothesis that there is a single latent variable, adaptive number knowledge, which influences performance on the ASP task for both whole numbers and rational numbers. This latent variable explains as much as two thirds of the variance in the ASP task scores (using the PCA model). The hypothesis that the latent variable is a joint form of adaptive number knowledge is strengthened by the fact that the factor loadings do not seem to be higher for one number type and lower for the other number type. In all three analyses, a rational number item (target 3) loaded most strongly onto the latent factor, followed by a whole number item (target 59), another rational number item (target ½), and another whole number item (target 38). Thus, whole number items did not load notably more strongly onto the latent factor than rational number items, nor vice versa.

This finding accords with the view of high adaptive number knowledge as a characteristic of students who also have high routine expertise in mathematics (McMullen et al., 2019). Those with high levels of whole number ANK would therefore be expected to also display general competence in whole number arithmetic, and competence in whole number arithmetic predicts competence in rational number arithmetic (see Section 1.5). A firm understanding of the concepts and procedures of rational number arithmetic, coupled with a well-connected mental model of whole number relations, would likely be a recipe for good ARNK. If a student already possesses a well-connected knowledge of the relations between whole numbers, it seems likely that they would be able to expand that framework to include rational numbers, once they had developed a good conceptual and procedural understanding of rational numbers. Therefore, it makes sense that – at least in older students who have already completed basic instruction on fractions and decimals – ANK and ARNK should behave as a single joint construct.

It is, of course, possible that both the whole number and rational number items are reflecting the influence of some other latent variable, such as general mathematical ability or arithmetic fluency. However, given that prior research has demonstrated that ARNK is distinct from routine arithmetic knowledge (McMullen et al., 2020) and that the ASP task is capable of differentiating between students who already have a high level of mathematical knowledge and skill (McMullen et al., 2019), it seems more likely that this latent factor is measuring something more than general mathematical skills and knowledge, which are usually measured in more routine ways.

Even if the identity of the latent factor cannot be verified, the EFA has at least demonstrated that whole number and rational number ASP items *do not measure different things*. Both types of items capture some aspect of a single underlying factor. Therefore, this EFA contributes to establishing convergent validity for ANK and ARNK. Future research may be able to establish discriminant validity by explicitly comparing performance on a combined whole and rational number ASP task to performance on measures of other mathematical skills.

However, one more possible explanation for the lack of divergence between ANK and ARNK must be considered. As noted in Chapter 2, the whole number items were sparse items (thus, more difficult) and the rational number items were dense items (thus, easier). This may have resulted in students generating relatively fewer whole number responses and more rational number responses than they would have if the question types had been comparable. As a result, the response patterns may have been relatively similar across all four input variables. However, in a scenario where the same type of task was given for both types of numbers, it seems much more likely that a difference would emerge, with distinct answer patterns on whole number and rational number items. This possibility should be investigated in future research.

To conclude, the exploratory factor analyses conducted in this chapter provided support for the hypothesis that adaptive whole number knowledge and adaptive rational number knowledge are *not* distinct, as both appear to be influenced by a single latent variable. This variable is likely to be adaptive number knowledge, although divergent validity can only be confirmed by further research. However, the decision to use sparse whole number items and dense rational number items may have obscured a real distinction between ANK and ARNK; this possibility should be investigated in future studies.

4. Individual Differences in ANK and ARNK

This chapter addresses the second research question, namely: *Are there quantitative and/or qualitative individual differences in high school students' ANK*[13] *and ARNK?* Again, in answering this question, only the data from the arithmetic sentence production task were considered. The intention of this chapter is to investigate how a sample of high school students performed on an established measure of A(R)NK, in the context of previous research on the subject. As all previous research has used the ASP task to measure adaptive number knowledge, it is appropriate to focus on the ASP task here.

The first section of this chapter gives an overview of the data by examining descriptive statistics from the ASP task. This provides a foundation for the investigations of individual differences that follow. The second section investigates quantitative differences in ANK and ARNK by means of one-way ANOVAs, independent samples t-tests and Kruskal-Wallis tests. The third section investigates whether qualitative differences in ANK and ARNK can be detected with the aid of a cluster analysis.

As noted in the introduction, the question of how A(R)NK develops over time is under-researched. Therefore, in many of the analyses that follow, special attention will be paid to the possibility of systematic differences in terms of age, school grade or mathematics module (as a proxy for mathematical experience).

4.1 Overview of the Data: Descriptive Statistics

As outlined in Chapter 2, there were four items in the ASP task. Two used exclusively whole numbers (target numbers 59 and 38), and two used mainly rational numbers (target numbers ½ and 3). Three variables were considered for each item in the ASP task. The first variable was the total number of correct solutions generated for that item. The second variable was the number of (correct) complex solutions generated for that item. As in prior research, the term *complex* is used to refer to solutions which combine additive operations (addition and/or

[13] As in Chapter 3, the abbreviation "ANK" is in this chapter to refer specifically to adaptive *whole* number knowledge. The full form "adaptive number knowledge" is used to refer to adaptive number knowledge in general, unless specified otherwise.

subtraction) with multiplicative operations (multiplication and/or division). The third variable was the number of (correct) solutions that included both fractions and decimals. This last variable was only relevant for the rational number ASP items.

The number of valid responses for each item ranged from $n = 420$ to $n = 425$. Given that the total sample size was $N = 447$, this meant that no more than 6% of the data was missing for any single item. As explained in Section 3.2.2, most missing data were due to illegible scans of the original test scripts, and can be considered to be missing completely at random.

4.1.1 Total Number of Correct Solutions

Descriptive statistics for the total number of correct solutions are given in Table 4.1. The minimum number of solutions for all items was zero, but it is interesting to note that the maximum number of solutions increased for each subsequent item, from 8 solutions for target number 59 to 14 solutions for target number 3. The reason that this should be the case is not immediately apparent, as the mean statistics did not increase monotonically from the first through to the fourth item. It may be due to the highest-scoring participants' getting used to the requirements of the novel ASP task, and thus speeding up as they proceeded through the items (or it may be due to random chance). It would be interesting to compare these results with the results of past ANK tests to see if the same pattern emerges. Unfortunately, this level of detail is not generally included in scientific articles, so it would need to be pursued specifically in future research. If there is a tendency for high performers to improve in the later tasks, this should be taken into consideration in the ordering of test items (for instance, by alternating between whole and rational items rather than clustering each type together) as well as in the analysis.

The mean statistics indicate that, in general, participants produced more correct solutions for the rational number items than for the whole number items. An average of 4.42 correct solutions were generated on the whole number items, while for the rational number items an average of 6.33 correct solutions were generated. This means that 43% more solutions were produced for rational number items than whole number items, which is remarkable given that students tend to find rational number arithmetic more challenging than whole number arithmetic. The hypothesis that students struggle more with rational numbers than whole

numbers is supported by decades of evidence (see e.g. Lortie-Forgues et al., 2015; Siegler et al., 2011), so it is unlikely that this research has uncovered a previously hidden population of students who prefer fractions to whole numbers. It is more likely that this finding results from the decision to include only sparse whole number items and only dense rational number items in this version of the ASP task. As noted in the conclusion to Chapter 3, this may have resulted in participants' generating relatively fewer whole number responses and more rational number responses rational number responses than they would have generated for comparable item types.

Table 4.1

Descriptive Statistics for the Total Number of Correct Solutions (ASP Task)

	N	Range	Min	Max	Mean	Std. Deviation
Total correct: Target 59	425	8	0	8	1.76	1.36
Total correct: Target 38	425	10	0	10	2.66	1.72
Total correct: Target ½	420	12	0	12	3.70	2.19
Total correct: Target 3	424	14	0	14	2.63	2.16
Valid N (listwise)	405					

The standard deviation scores were also noticeably larger for the rational number items, indicating that the variance within the sample was larger for the rational number items. This is unsurprising: since the rational number items were dense, a student who was skilled in rational number arithmetic should have been able to generate a large number of solutions (larger than on the whole number items); but since many students struggle with rational numbers it is not surprising that weaker students may have achieved very low scores on these items. Therefore, this is a natural consequence of the individual variation one would expect to find on such an item.

4.1.2 Number of Complex Solutions

Descriptive statistics for the number of complex solutions are given in Table 4.2. As would be expected, the average number of complex solutions on the sparse whole number items was high relative to the number of total solutions on those items. In fact, these results indicate that on average, 81.7% of solutions generated for target 59 and 81.2% of solutions generated for target

38 were complex. By contrast, an extremely small number of solutions generated for the rational number items were complex: only 7% of all solutions for target ½ and 10.8% of all solutions for target 3. This is likely to be primarily a function of the question type: it is not necessary to generate complex solutions on a dense item where many simple solutions are readily apparent. However, it may also be reflective of a reluctance or inability to perform more complex operations with rational numbers. This hypothesis cannot be tested without comparable data, emphasising again the importance of using both sparse and dense items with both number types in future research. By comparing the mean to the standard deviations and the range, it is clear that the distribution of complex solutions is skewed to the right. This means that a small number of respondents generated a relatively high number of complex responses even on the rational number items. This is consistent with previous research, which has found consistently that a small group of high achievers is able to produce many more complex solutions that their peers (McMullen et al., 2016, 2017, 2019), and is indicative of individual differences in the data.

Table 4.2

Descriptive Statistics for the Number of Complex Solutions (ASP Task)

	N	Range	Min	Max	Mean	Std. Deviation
Complex solutions: Target 59	425	6	0	6	1.44	1.16
Complex solutions: Target 38	425	7	0	7	2.16	1.52
Complex solutions: Target ½	420	4	0	4	0.26	0.69
Complex solutions: Target 3	424	6	0	6	0.29	0.73
Valid N (listwise)	405					

4.1.3 Number of Solutions Combining Fractions and Decimals

Descriptive statistics for the number solutions combining fraction and decimal notation are given in Table 4.3. Each rational number task contained four fractional values among the given numbers, two in fraction form and the same two in decimal form (for instance: ¼, ¾, 0.25, 0.75 and 2 were the given numbers for target ½ – see Chapter 2 for further detail). This meant that in each rational number item, it would be possible to generate multiple versions of a solution by swapping decimal and fraction representations. For instance, ¼ + ¼ , 0.25 + ¼, ¼ + 0.25 and 0.25 + 0.25 would all be valid and distinct solutions. Therefore, one might expect solutions

like ¼ + 0.25 to be about as common as solutions like ¼ + ¼. However, they seemed to be far less common. An average of 0.62 cross-notational solutions were generated for target number ½, which is only 16.7% of all solutions for that item. For target number 3, the proportion is only marginally higher at 20.1% of all solutions.

Table 4.3

Descriptive Statistics for the Number of Solutions Combining Fractions & Decimals (ASP Task)

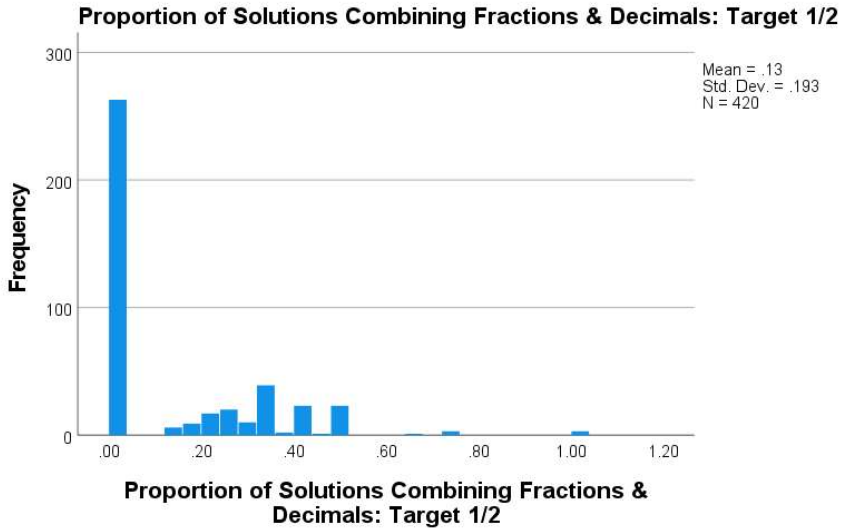
	N	Range	Min	Max	Mean	Std. Deviation
Solutions combining fractions & decimals: Target ½	420	4	0	4	0.62	0.92
Solutions combining fractions & decimals: Target 3	424	6	0	6	0.53	0.87
Valid N (listwise)	405					

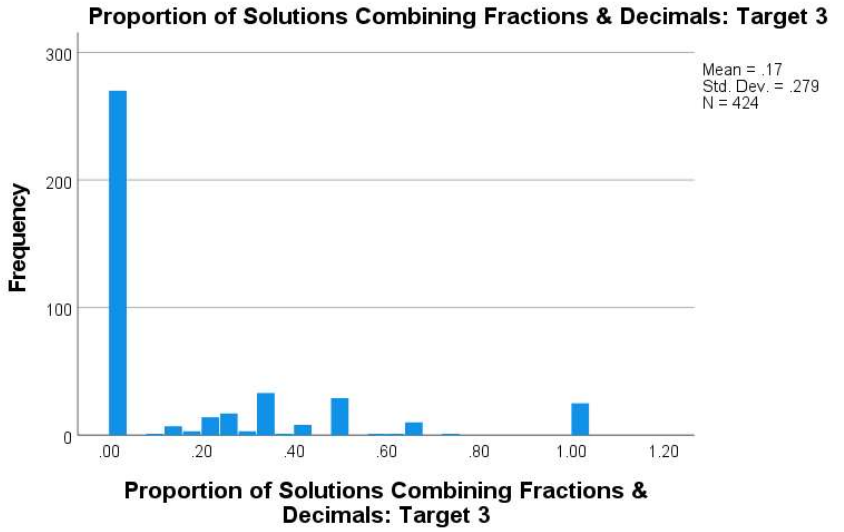
The number of *students* who generated at least one solution combining fractions and decimals is considerably larger at 222 (54% of participants). However, as Figure 4.1 shows, this means there were still 192 participants (46%) who did not generate any cross-notational solutions. Furthermore, many of the participants who generated at least one cross-notational solution did so only on one item: only 157 participants (37%) produced cross-notational solutions for target number ½, and only 154 participants (36%) produced cross-notational solutions for target number 3. The histograms in Figure 4.1 also confirm that, for most participants, cross-notational solutions made up a small proportion of their solutions.

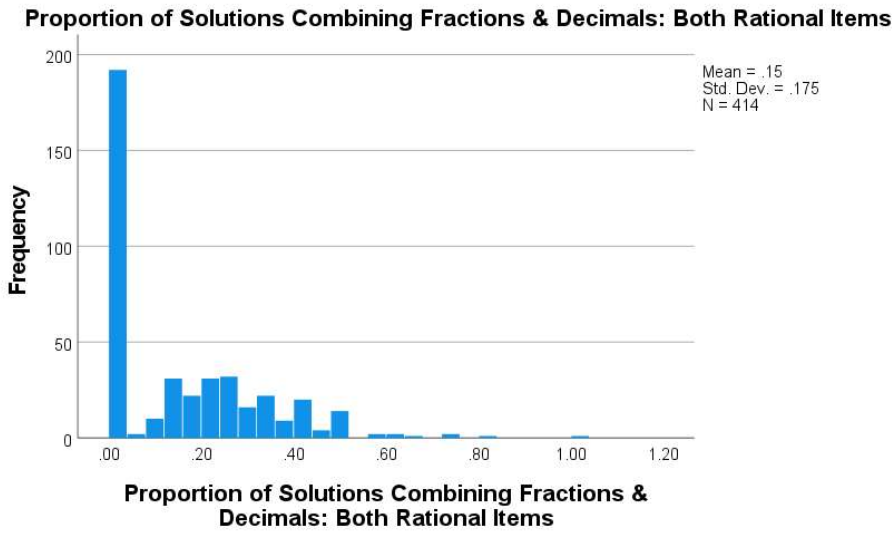
This finding is more telling than the similar result regarding complex solutions. That is because the form of the dense items made simple solutions much easier and faster to generate than complex solutions – but no such task-based constraint would favour the use of single-notation solutions over cross-notational solutions. Therefore, the paucity of cross-notational solutions likely indicates a lack of integration between fraction and decimal understanding among many of the participants. This would be consistent with previous findings that even high school students often struggle to grasp the connection between fractions and decimal numbers. For instance, Vamvakoussi and Vosniadou (2010) found that seventh-, ninth- and eleventh-graders

Figure 4.1

Histograms of the Proportion of Solutions Combining Fractions and Decimals (ASP Task)







tended to believe that there could be only fractions (but not decimals) in the interval between two fractions, and only decimals (but not fractions) in the interval between two decimals.

Failing to grasp the connection between fractions and decimals is intimately connected to adaptive rational number knowledge. By definition, students with a high level of ARNK have a well-connected understanding of rational numbers and their relations to each other – therefore, by definition, one must understand the connections between different types of rational numbers (like fractions and decimals) in order to reach the highest levels of ARNK. The fact that many students did not integrate fractions and decimals in their solutions suggests that they lacked a complete conceptual understanding of the nature of rational numbers, which is fundamental to well-developed ARNK. This interpretation is supported by the fact that the number of cross-notational solutions generated is strongly correlated with the total number of solutions generated on rational number items (Pearson's $r = 0.61$, $p < 0.001$), and indeed on all items in the ASP task (Pearson's $r = 0.60$, $p < 0.001$).

An alternative explanation might be that students did not realise that they were allowed to combine fraction and decimal notation in this task, but that would amount to the same problem – not being aware that fractions and decimals are simply alternate notations for the same kind of number, which can be used interchangeably. By the time the participants answered the rational number items, they had already answered a practice item and two whole number test items, and so should have been comfortable with the idea that *any* of the given numbers could be combined to reach the target number. Nonetheless, an example or practice item combining fractions and decimals may have better clarified the "rules of the game" for the rational number items, and could be considered for inclusion in future versions of this test.

Therefore, these data suggest that there are individual differences between students who understand and apply the connection between decimals and fractions, and those who do not. Importantly, one need not utilise *exclusively* cross-notational solutions in order to have strong ARNK – and indeed, restricting oneself to cross-notational solutions would be likely to limit the number of solutions one could generate in the given time, since easy options like ¼ + ¼ and 0.25 + 0.25 would be off limits, even though they would probably spring to mind when writing down solutions like ¼ + 0.25. In fact, it is interesting to observe that out of the 25 students who generated exclusively cross-notational solutions for target 3, none of them

generated more than three solutions for that item, and most generated less.[14,15] This places their total scores, at best, only slightly above the mean number of solutions, and in most cases somewhat below the mean. Therefore, it may be the case that an exclusive focus on cross-notational answers was a maladaptive strategy choice for a task where the explicit aim was to generate as many solutions as possible. However, whether it was a conscious strategy choice cannot be determined with the existing data; that would require a different research design including interviews or some other way of accessing the participants' thought process.

4.2 Quantitative Differences in A(R)NK Between Groups

4.2.1 Differences Between Age Groups

Total Number of Correct Solutions

Before running a one-way ANOVA, the suitability of the data for that procedure was considered. As mentioned in Chapter 2, the age breakdown of the sample was as follows: one 13-year-old, 119 14-year-olds, 169 15-year-olds, 86 16-year-olds, 43 17-year-olds and 13 18-year-olds. The single 13-year-old was removed from the data set for the purposes of this analysis, since a single case cannot constitute a group.

Furthermore, the normality of the 17-year-old and 18-year-old groups was considered, since they were small enough that the Central Limit Theorem might not apply. In both groups, an inspection of histograms showed that the distribution of solutions for target 59 was clearly non-normal (monotonic decreasing). It was therefore decided to use the subtotals for whole number items and rational number items instead, since the subtotal distributions were reasonably close to normal. The subtotals for whole and rational items were also used for all subsequent between-groups analyses on the number of correct solutions, so that the findings would be comparable.

[14] Twenty out of the 25 generated one solution, three generated two solutions, and two generated three solutions.
[15] It is also interesting to observe that only three students generated exclusively cross-notational solutions for target ½, although it is unclear why this number is so much smaller than for target 3. These three students generated one, one and two solutions for that item.

A one-way ANOVA was then run to test the effect of age on A(R)NK scores. The tests revealed that age had no significant effect on either the whole number scores, $F(4,400) = 0.4$, $p = 0.81$, or the rational number scores, $F(4,397) = 1.32$, $p = 0.26$.

Number of Complex Solutions

Because so many of the responses for whole number ASP items, and so few of the responses for rational number ASP items, were complex, it was decided to analyse the total number of complex solutions across all four items. This decision applies to all between-groups analyses of complex solutions. Again, the single 13-year-old was removed from the data set, and normality checks were conducted in the 17-year-old and 18-year-old subgroups. The distribution in the 17-year-old subgroup was reasonably normal, with skewness (1.13) and kurtosis (1.42) figures confirming the mild rightwards skew in a visual inspection of the histogram. The data in the 18-year-old subgroup were in fact more normally distributed, with skewness (0.74) and kurtosis (−0.88) figures indicating no significant departure from normality.

A one-way ANOVA was then run to test the effect of age on the number of complex solutions generated. The test showed that age had no significant effect on complex solution generation, $F(4,388) = 0.99$, $p = 0.42$.

Number of Solutions Combining Fractions and Decimals

It was decided to use the total number of cross-notational solutions across both rational number item tasks as the dependent variable in this ANOVA. Because the number of cross-notational solutions is small, combining the two scores was considered to be the best strategy in order to maximise normality and inter-group variance. The total number of cross-notational responses will be used as the dependent variable for all between-groups comparisons looking at cross-notation.

As above, the single 13-year-old was removed from the data set, and normality checks were conducted in the 17-year-old and 18-year-old subgroups. Both groups showed a non-normal distribution. The 17-year-old group in particular showed a very strong rightwards skew, with heavy tails (kurtosis = 5.15). The 18-year-old group showed a rightwards skew but had less

heavy tails. The presence of heavy tails in the 17-year-old group is important, because while the Central Limit Theorem usually guarantees a normal sampling distribution in a sample of 30 or more, in a heavy-tailed distribution the required sample size must be much larger (sometimes as much as 160) before a normal sampling distribution can be assumed (Field, 2018). Since the assumption of normality was violated, it was decided to compare groups with the non-parametric Kruskal-Wallis test. The Kruskal-Wallis test revealed that the number of cross-notational solutions was not significantly influenced by age, $H(4) = 6.37$, $p = 0.17$.

4.2.2 Differences Between School Grades

Total Number of Correct Solutions

Before running a one-way ANOVA, the suitability of the data for that procedure was considered. As mentioned in Chapter 2, the grade breakdown of the sample was as follows: 236 ninth-graders, 114 tenth-graders, 62 eleventh-graders and 30 twelfth-graders. Since the twelfth-grade group was small enough that the Central Limit Theorem might not apply, normality of the relevant variables was inspected within the group. Visual inspection of the histograms indicated that the numbers of whole and rational solutions were pleasingly normally distributed; this was confirmed by statistics for skewness that were under 0.35 in both cases.

A one-way ANOVA was then run to test the effect of school grade on A(R)NK scores. The tests revealed a significant effect of school grade on both whole number scores, $F(3,411) = 4.33$, $p = 0.005$, and rational number scores, $F(3,406) = 7.90$, $p < 0.001$.

Levene's test indicated that the variances did not differ significantly between school grades for whole number items, $F(3,411) = 1.92$, $p = 0.13$, or rational number items, $F(3,406) = 0.19$, $p = 0.90$. Therefore, the Hochberg GT2 post hoc test was used, since Field (2018) recommends this test in circumstances where the variances are homogeneous but the sample sizes are very different.

The post hoc comparisons indicated that among the whole number ASP items, the mean number of solutions was significantly lower in the eleventh-grade group (M = 3.59, SD = 3.25) than in the twelfth-grade group (M = 5.76, SD = 3.02). No other significant differences were

detected in the whole number scores. Among the rational number ASP items, the mean number of solutions in the eleventh-grade group (M = 4.25, SD = 3.94) was found to be significantly lower than the other three groups (ninth-grade M = 6.47, SD = 3.84; tenth-grade M = 7.11, SD = 3.61; twelfth-grade M = 7.32, SD = 4.11). No significant differences were detected between any other grades.

Number of Complex Solutions

As above, the distribution of the relevant variable (in this case, the total number of complex solutions generated in the ASP task) was checked in the small twelfth-grade group. An inspection of the histogram, together with the skewness and kurtosis figures, did not reveal any major deviations from normality. Therefore, it was decided to proceed with a one-way ANOVA to test the effect of school grade on number of complex solutions.

The ANOVA revealed a significant effect of school grade on the number of complex solutions, $F(3,397) = 5.29$, $p = 0.001$. Levene's test indicated that the variances did not differ significantly between school grades (based on the median value, as recommended by Field, 2018), $F(3,397) = 2.36$, $p = 0.07$. Therefore, the Hochberg GT2 post hoc test was used. The post hoc comparisons indicated that the mean number of complex solutions was significantly lower in the eleventh-grade group (M = 3.13, SD = 3.53) compared to the twelfth-grade group (M =5.71, SD = 3.45), but no other significant differences were detected.

Number of Solutions Combining Fractions and Decimals

As above, the distribution of the relevant variable (the total number of cross-notational solutions generated in the ASP task) was checked in the small twelfth-grade group. This group shows a strongly skewed distribution, with a long right-hand tail and a much higher frequency of 0 cross-notational solutions than any other number of solutions. Kurtosis was therefore also fairly high at 2.45. To avoid possible problems stemming from the violation of the assumption of normality in a one-way ANOVA, it was decided to compare groups with the non-parametric Kruskal-Wallis test. The Kruskal-Wallis test revealed that the number of cross-notational solutions was affected by school grade, $H(3) = 20.61$, $p < 0.001$. Pairwise comparisons with adjusted *p*-values showed that the eleventh-grade group generated significantly fewer cross-

notational solutions than students in the other three school grades. No other differences were detected. The pairwise comparisons used the Bonferroni correction for multiple tests.

4.2.3 Differences Between Mathematics Modules

Total Number of Correct Solutions

Before conducting an ANOVA, the normality of the relevant variables was examined for the three smallest groups: Algebra 1 Honors ($n = 42$), Geometry ($n = 21$) and AP Statistics ($n = 29$). No significant departures from normality were observed.

A one-way ANOVA was then run to test the effect of mathematics module on A(R)NK scores. The test revealed a significant effect of mathematics module on both whole number scores, $F(5,413) = 18.66$, $p < 0.001$, and rational number scores, $F(5,408) = 21.75$, $p < 0.001$.

Levene's test indicated that the variances did not differ significantly between mathematics modules for the whole number items, $F(5,413) = 2.28$, $p = 0.05$, or for the rational number items, $F(5,408) = 1.41$, $p = 0.22$. Therefore, the Hochberg GT2 post hoc test was used.

The post hoc comparisons indicated that among the whole number ASP items, the mean scores differed as follows (refer to Table 4.4 for specific values):

- AP Statistics was significantly higher than all other modules.
- Geometry Honors was significantly higher than Algebra 1 Honors, Algebra 2 and Geometry.
- Algebra 2 Honors was significantly higher than Algebra 1 Honors and Geometry.

Among the rational number ASP items, the post hoc comparisons indicated exactly the same differences as in the whole number items, with one addition: the mean score for Algebra 2 Honors was significantly higher than for Algebra 2.

Table 4.4

Descriptive Statistics for One-Way ANOVA Between Total Correct Solutions (ASP Task) & Mathematics Module

	Correct responses on whole number ASP items				Correct responses on rational number ASP items			
	N	**Mean**	**Std. Deviation**	**Std. Error**	**N**	**Mean**	**Std. Deviation**	**Std. Error**
Algebra 1 Honors	42	2.86	1.93	0.30	42	3.69	3.47	0.54
Algebra 2	54	3.44	2.08	0.28	50	4.26	3.12	0.44
Algebra 2 Honors	124	4.61	2.76	0.25	121	7.28	3.90	0.35
AP Statistics	29	7.69	3.06	0.57	29	9.97	3.67	0.68
Geometry	19	2.26	2.10	0.48	20	2.45	2.16	0.48
Geometry Honors	151	4.70	2.39	0.19	152	6.82	3.41	0.28
TOTAL	**419**	**4.42**	**2.72**	**0.13**	**414**	**6.34**	**3.92**	**0.19**

Number of Complex Solutions

The distribution of the number of complex solutions was again investigated for the three smallest mathematics modules, Algebra 1 Honors, Geometry and AP Statistics. No major diversions from normality were observed. A one-way ANOVA was then conducted to test the effect of mathematics module on the number of complex solutions generated. The test revealed a significant effect of mathematics module on the number of complex solutions, $F(5,399) = 18.85$, $p < 0.001$.

Levene's test indicated that the assumption of homogeneity of variances had been violated, $F(5,399) = 4.86$, $p < 0.001$. It was therefore decided to use the Games-Howell post hoc procedure, which Field (2018) recommends in cases of unequal variance.

The post hoc comparisons indicated that, the mean number of complex solutions differed as follows (please refer to Table 4.5 for specific values):

- AP Statistics was significantly higher than all other modules.
- Geometry Honors was significantly higher than Algebra 1 Honors, Algebra 2 and Geometry.

- Algebra 2 Honors was significantly higher than Algebra 1 Honors, Algebra 2 and Geometry.

Table 4.5

Descriptive Statistics for One-Way ANOVA Between Total Complex Solutions (ASP Task) & Mathematics Module

	Total number of complex solutions (all ASP items)			
	N	**Mean**	**Std. Deviation**	**Std. Error**
Algebra 1 Honors	42	2.60	2.06	0.32
Algebra 2	50	2.86	1.76	0.25
Algebra 2 Honors	120	4.38	3.02	0.28
AP Statistics	29	7.59	3.83	0.71
Geometry	18	1.78	1.77	0.42
Geometry Honors	146	4.47	2.46	0.20
TOTAL	**405**	**4.15**	**2.90**	**0.14**

Number of Solutions Combining Fractions and Decimals

The distribution of the number of cross-notational solutions was investigated for the three smallest mathematics modules, Algebra 1 Honors, Geometry and AP Statistics. Although the distribution for the AP Statistics group was relatively normal, the Geometry group and (especially) the Algebra 1 Honors group showed major departures from normality. Zero cross-notational solutions (the minimum possible score) was by far the most commonly-occurring count in both groups, accounting for 13 out of 20 values in the Geometry group (65%), and 28 of the 42 values in the Algebra 1 Honors group (67%). This resulted in a strong right skew, especially in the Algebra 1 Honors group. It was decided to compare groups with the non-parametric Kruskal-Wallis test in order to avoid possible problems stemming from using non-normal data in a one-way ANOVA.

The Kruskal-Wallis test revealed that the distribution of cross-notational solutions differed significantly between mathematics modules, $H(5) = 43.47$, $p < 0.001$. Pairwise comparisons with adjusted p-values, using the Bonferroni correction for multiple tests, revealed that the number of cross-notational solutions differed as follows (refer to Table 4.6 for specific values):

- AP Statistics was significantly higher than Algebra 1 Honors ($p = 0.000$), Algebra 2 ($p = 0.000$), and Geometry ($p = 0.001$). The difference between the AP Statistics group and the Geometry Honors group was only borderline significant ($p = 0.046$), and the difference between the AP Statistics group and the Algebra 2 Honors group was not significant ($p = 0.11$).
- In addition, Algebra 2 was significantly lower than Geometry Honors ($p = 0.001$) and Algebra 2 Honors ($p = 0.000$).

Table 4.6

Descriptive Statistics for Multi-Notational Solutions (ASP Task) by Mathematics Module

	N	Range	Min	Max	Mean	Std. Deviation
Algebra 1 Honors	42	4	0	4	0.74	1.21
Algebra 2	50	3	0	3	0.42	0.81
Algebra 2 Honors	121	7	0	7	1.33	1.48
AP Statistics	29	9	0	9	2.52	2.34
Geometry	20	2	0	2	0.45	0.69
Geometry Honors	152	6	0	6	1.20	1.29
TOTAL	**414**	**9**	**0**	**9**	**1.15**	**1.45**

4.2.4 Differences Between Genders

Total Number of Correct Solutions

An independent samples t-test was conducted to evaluate the effect of gender on the total number of correct solutions obtained in the ASP task.[16] Once missing values and the single non-binary response were excluded, the sample included 198 females and 213 males.

On average, males (M = 4.99, SD = 2.85) scored significantly higher than females (M = 3.89, SD = 2.47) on whole number items, $t(407.15) = -4.18$, $p < 0.001$, Cohen's $d = 0.41$[17]. Since

[16] This test was initially conducted on a whim and had no theoretical basis, but since the solutions were unexpectedly interesting, it has been included in this chapter.

[17] All calculations of Cohen's *d* reported in this chapter use pooled standard deviation as the denominator.

$d = 0.2$ would be a small effect size and $d = 0.5$ would be a medium effect size (Field, 2018), the effect of gender on whole number ASP scores can be regarded as moderate.

Males (M = 7.23, SD = 4.22) also scored significantly higher than females (M = 5.48, SD = 3.37) on rational number items, $t(393.31) = -4.63$, $p < 0.001$, Cohen's $d = 0.46$. This suggests that the effect of gender on rational number ASP scores can also be regarded as moderate, resulting in a score difference of 0.46 standard deviations on average.

Section 4.2.3 suggests that the mathematics module in which a student is enrolled has a significant effect on the number of solutions they produce. It was therefore decided to conduct a chi-square analysis to establish whether male participants were more likely than females to be enrolled in higher-level classes. However, the chi-square test was not significant, $\chi^2(5,429) = 3.04, p = 0.70$, which indicates that there is no significant association between gender and mathematics module enrolment. This means that gender and mathematics module are both related to A(R)NK, but in distinct and separate ways.

Number of Complex Solutions

An independent samples t-test was conducted to evaluate the effect of gender on the number of complex solutions generated in the ASP task. Once missing values and the single non-binary response were excluded, the sample included 193 females and 204 males.

On average, males (M = 4.72, SD = 3.02) scored significantly higher than females (M = 3.62, SD = 2.68), $t(395) = -3.82$, $p < 0.001$, Cohen's d=0.38. This result suggests that the effect of gender on the generation of complex solutions is small to moderate.

Number of Solutions Combining Fractions and Decimals

An independent samples t-test was conducted to evaluate the effect of gender on the number of cross-notational solutions generated in the ASP task. Once missing values and the single non-binary response were excluded, the sample included 197 females and 209 males.

On average, males (M = 1.50, SD = 1.66) scored significantly higher than females (M = 0.81, SD = 1.12), $t(367.44) = -4.98$, $p < 0.001$, Cohen's d=0.49. This result suggests that the effect of gender on the generation of cross-notational solutions is moderate.

4.2.5 Discussion of Between-Group Differences

Age and School Grade

The results in Section 4.2.1 reveal that students' age did not affect the total number of correct solutions, the number of complex solutions, or the number of cross-notational solutions generated. This is interesting, because previous studies with elementary and lower-secondary school students have found that age is positively related to membership of adaptive profiles (McMullen et al., 2016, 2017) and that overall scores on the ASP task tend to increase with age (as outlined in Section 1.6). The present results suggest that perhaps A(R)NK develops rapidly in childhood and pre-adolescence, as children develop cognitively and learn about whole numbers and (later) basic rational numbers in increasing detail and complexity at school, but levels off in upper secondary school. This might be because the conceptual and procedural framework of real numbers is largely established by the time students reach upper secondary school. It is assumed, but seldom explicitly taught in the last years of school mathematics instruction.

Another possibility is that this finding is the result of country-specific factors. The studies which uncovered age-related differences in ANK were all conducted in Finland, while the only study to measure adaptive number knowledge in multiple grades conducted outside of Finland (a study of ARNK in US 7th and 8th graders, McMullen et al., 2020) did not find any differences in ARNK between grades. This result might reflect the levelling-off hypothesised in the previous paragraph, or it might reflect a flatter A(R)NK profile among American youngsters, which would also be consistent with this paper's findings. The only way to know for sure is to conduct and compare studies of similar-aged students in different countries.

One feature of the American system that might well explain these observations is the modular system of high school mathematics courses. In Finnish basic education (Grades 1-9), students follow a set national curriculum, which means that there is a clear progression of mathematics

content from year to year. All students are expected to start school at the same age and repeating grades is very rare (Finnish National Agency for Education, 2018). This means that as students grow older, they necessarily participate in more advanced mathematics classes. The age effect that is apparent in prior research could therefore be due either to age or to mathematical experience. In American high schools (and to a lesser extent in American middle schools), by contrast, the mathematics curriculum is modular and much more flexible. Although students are typically required to take some mathematics modules in order to graduate from high school, they are able to choose relatively freely between more and less advanced modules. For example, the high school at which the present research was conducted offered 17 different year-long mathematics modules, which varied in both depth and scope. Mathematically-inclined students can enrol in more demanding modules (such as Honors courses) and might be able to take more than one module each year, while other students can take fewer and easier modules. This means that age does not correlate neatly with the amount of mathematical experience a student has. This opens the door to two further possible interpretations of the data, in addition to levelling-off of A(R)NK among adolescents. The first option is that age has never had any bearing on students' A(R)NK levels and the variance seen in prior studies was due entirely to higher levels of mathematical experience amongst older learners. The second option is that while age does affect A(R)NK, mathematical experience and expertise affects it more, and in this sample the two effects are working in opposite directions, thus obscuring the effects of age. It can be seen in Table 4.7 that in this sample, younger students are more likely to be enrolled in demanding Honors and AP courses than older students. It seems plausible that students who enrol in more demanding mathematics modules would have greater mathematical exposure and expertise than students who do not. Thus, this might cancel out any age-related effects.

School grade is highly correlated with age in this sample (Pearson's $r = 0.84$, $p < 0.001$). One might therefore expect school grade to be a similarly poor predictor of performance on the ASP task. Inter-group comparisons indeed failed to reveal any significant differences between school grades, except in the case of 11^{th} graders, who generally scored lower than their peers in other grades. Closer examination of the eleventh-grade subsample suggests that students in this grade seem to differ systematically from students in other grades, being less likely to enrol in Honors or AP classes (see Table 4.7). One possible interpretation is that these 11^{th} graders have not engaged with mathematics of the same depth and complexity when compared with their peers in other grades, which might have had a negative impact on their A(R)NK

development. A slightly different interpretation is that these students may have weaker mathematical skills than their other-grade peers, which have *caused* them to register for less demanding courses, and which would also reflect in poorer performance on the ASP task.

Table 4.7

Enrolment in Honors and Advanced Placement Modules by School Grade and Age

Grade	Honors/AP	Other	Total	Age	Honors/AP	Other	Total
9	233 (98.7%)	3 (1.3%)	236 (100%)	14	118 (99.2%)	1 (0.8%)	119 (100%)
10	78 (68.4%)	36 (31.6%)	114 (100%)	15	145 (85.8%)	24 (14.2%)	169 (100%)
11	36 (58.1%)	26 (41.9%)	62 (100%)	16	59 (68.6%)	27 (31.4%)	86 (100%)
12	21 (70%)	9 (30%)	30 (100%)	17	27 (62.8%)	16 (37.2%)	43 (100%)
				18	8 (61.5%)	5 (38.5%)	13 (100%)
ALL GRADES	**368 (83.3%)**	**74 (16.7%)**	**442 (100%)**	**ALL AGES**	**354 (82.3%)**	**76 (17.6%)**	**430 (100%)**

Importantly, it seems improbable that 11th graders in the general American high school population would be much less likely than those in other grades to take advanced mathematical courses. This means that the present sample may not be fully representative of the broader population; thus, caution should be applied when generalising the findings of this study. Precise figures on the number of students taking Honors courses in high school are not readily available, largely because the definition of an Honors course varies from school to school. Figures on AP courses are more easily available, as AP courses are centrally accredited by an organisation called the College Board, which also administers examinations for those courses. A total of 450 484 twelfth-grade students wrote the AP examinations for Calculus AB, Calculus BC and/or Statistics in 2019, the most recent year of data that was unaffected by the COVID-19 pandemic (College Board, 2019).[18] The total twelfth-grade population of the USA was approximately 4.5 million in 2019 (Data Commons, 2022), suggesting that at most

[18] The number of unique candidates is probably somewhat lower, as some candidates may have written an examination for more than one AP subject.

approximately 10% of students wrote examinations in AP mathematics. The number of students who enrol for any form of Honors or high-performance mathematics course during the course of their high school career is probably significantly higher (especially since one must pay per subject to write the AP examinations). However, this number is also known to vary widely from school to school, with some schools expecting virtually all students to take some Honors courses, and other schools not offering Honors courses at all. At the school where the present data were collected, taking Honors and AP mathematics classes was quite common, with perhaps around 70% of students taking such classes.[19] The results of this research should therefore be interpreted in the context in which it was collected. It seems likely that the results of the ninth- and tenth-grade groups which dominate this data set are not representative of the broader US high school population, although they are probably more representative of college-track high school students, who tend to take more advanced mathematics courses in preparation for college or university admissions, particularly if they wish to apply to mathematics- or science-focused programmes.

Mathematics Modules

Enrolment in more advanced mathematics modules showed a clear association with better performance on all three metrics: total number of correct solutions, complex solutions, and cross-notational solutions. Students in the college-credit AP Statistics module performed significantly better than students in any other module, and students enrolled in Algebra 2 Honors and Geometry Honors (the next most advanced modules) tended to outperform their peers in the lower-level Algebra 2, Geometry and Algebra 1 Honors modules. This is a new contribution to the A(R)NK literature, since earlier studies have generally focused on younger students, who are more likely to attend a single mathematics class with the rest of their age cohort.

The direction of the relationship is unknown. It might be that participation in higher-level mathematics classes improves A(R)NK by exposing students to demanding tasks which require them to apply and integrate their existing numerical knowledge in new ways. However, it might also be the case that high-achieving mathematics students enrol for higher-level mathematics

[19] The figure of 70% is an estimate based on conversations with the researcher who collected the data and is familiar with the school.

classes, and coincidentally these are also the students who have high levels of A(R)NK. Since it is already known that A(R)NK is a distinguishing characteristic amongst those with high levels of routine mathematical expertise (McMullen et al., 2019, 2020), this alternative explanation is plausible. A longitudinal study would be required to clarify the directionality (or bidirectionality) of the causal relationship.

Gender

It appears that male students produce significantly more solutions, significantly more complex solutions, and significantly more cross-notational solutions in the ASP task than their female counterparts. The only other variable in these analyses that showed a consistent association with performance on the ASP task was mathematics module enrolment. However, the chi-square analysis indicated that mathematics module enrolment was independent of gender, so this seems to be a robust finding in its own right.

To the best of the author's knowledge, this is a completely novel line of enquiry in adaptive number knowledge research. No previous studies have investigated the relationship between gender and A(R)NK, but it seems that it may be an important topic for further investigation. It is possible that the gender gap detected in this data set is reflective of higher mathematical adaptivity in general in (American) boys as compared to (American) girls. This would be consistent with the results of international benchmarking tests. American boys outperform American girls in PISA tests, which assess how well students are able to apply their mathematical knowledge and skills to everyday situations (i.e. the PISA test requires adaptive expertise),[20] yet there is no gender difference between American youngsters on the TIMSS assessments, which test more routine factual and procedural knowledge (ACER, n.d.; OECD, 2019b; TIMSS & PIRLS International Study Center, 2021). Furthermore, it is consistent with findings in developmental psychology that adolescent boys tend to outperform adolescent girls on mathematics tests that are less related to what is taught in school (Gallagher et al., 2000; Ganley, 2018), although girls perform at least as well as boys on routine tests (Ganley, 2018; Hyde et al., 2008). These differences appear earlier in higher-achieving students (Cimpian et al., 2016), who are the students one would expect to have higher A(R)NK, and are more

[20] Recall that A(R)NK is a component of adaptive expertise. This means that understanding individual differences in A(R)NK may help to explain *why* differences in performance emerge on tests like PISA.

pronounced in high-school students (Ganley, 2018; Hyde et al., 1990). Investigating the relationship between A(R)NK and gender, and specifically whether it holds across ages, countries, and within-country instructional contexts, could contribute to a better understanding of how adaptive expertise in mathematics develops.

4.3 Investigating Qualitative Differences with a Cluster Analysis

Previous research has found significant individual differences in performance on the ASP task through person-centred analyses (see Chapter 1). It was therefore decided to use a cluster analysis to investigate whether the present sample contained groupings that differed qualitatively in terms of their adaptive number knowledge. According to Niemivirta et al. (2019), a person-centred statistical procedure can always find groups in the data, so it is important that the execution and interpretation of a person-centred approach is informed by theory. The theoretical foundation of the proposed cluster analysis lies in both the literature on adaptive number knowledge, and the literature on rational number knowledge.

Prior research on adaptive number has shown that qualitative differences can often be detected between students who predominantly use complex solutions, and those who predominantly use simple solutions (McMullen et al., 2016, 2017, 2019). Furthermore, research in the field of rational number understanding suggests that the characteristics of rational numbers are so distinct from those of whole numbers that they require new models of thinking and conceptual understanding (see Section 1.5). Although Chapter 3 has shown that the total number of solutions generated for whole number ASP items is strongly associated with the total number of solutions generated for rational number ASP items, there still remains the possibility that differences will emerge when the nature of those solutions is analysed. Recall from Section 4.1.2 that the distribution of complex solutions was skewed to the right: this demonstrated that a small group of students was producing many more complex solutions than their peers, even on the rational number items. It might be that these students fall into a particularly high-scoring or otherwise adaptive profile.

Therefore, the topic of interest is whether different groups of students use simple and complex solutions in different ways on whole and rational number items. The input variables for the cluster analysis will be: (1) simple solutions on whole number items, (2) complex solutions on

whole number items, (3) simple solutions on rational number items, (4) complex solutions on rational number items. As a reminder, a complex solution is one which contains both additive and multiplicative operations. A simple solution is one which contains either additive or multiplicative operations, but not both. These definitions are complementary, so between them they encompass all possible correct solutions.

4.3.1 The Chosen Clustering Method: TwoStep Cluster Analysis

Because this cluster analysis is exploratory in nature, it was decided to use TwoStep Cluster Analysis. This method is advantageous in a situation where the underlying cluster structure is unknown because it is able to automatically select the "best" number of clusters based on the data and provides measures of fit that assist in choosing between cluster models (IBM Corporation, 2021a). Gelbard et al. (2007) have found that the TwoStep clustering method is superior to hierarchical algorithms, which would be the other option for open-ended clustering in SPSS.[21] The validity of the TwoStep clustering technique has also been confirmed by Benassi et al. (2020), who found that it produced reliable solutions which were similar to those produced by latent class analysis, a sophisticated method which is considered to have significant advantages over traditional methods of clustering (Hickendorff et al., 2018).

The TwoStep clustering method first divides the data set into sub-clusters using a log-linear or Euclidean distance measure, and then combines the sub-clusters into groups, using a probabilistic model (similar to latent class analysis) to choose the optimal subgroup model (Gelbard et al., 2007; Kent et al., 2014). It is therefore a hybrid model which combines elements of traditional distance-based clustering algorithms with the probabilistic approach of latent class analysis (Kent et al., 2014).

4.3.2 The Cluster Analysis: Choosing a Model

A TwoStep cluster analysis was run on the four indicator variables, using a log-likelihood distance measure, and asking SPSS to determine the number of clusters automatically.

[21] The third type of cluster analysis supported by SPSS, K-Means Cluster Analysis, requires that the number of clusters be predetermined and therefore is less straightforward in a purely exploratory context. Furthermore, Gelbard et al. (2007) found that TwoStep analysis performs just as well as K-Means clustering, meaning that the easier choice does not result in a loss of functionality.

Akaike's Information Criterion (AIC) was selected as the clustering criterion. The automatic procedure generated a two-cluster solution. The fit of this two-cluster solution was deemed by SPSS to be "fair" but not "good", having a silhouette coefficient of 0.4.[22] Since this is the optimal clustering solution found by SPSS, this suggests that no better clustering solution exists in the given data (although solutions of similar quality do exist, as seen below). The fact that the best solution is merely a "fair" solution may suggest that the data is not ideally-suited to clustering, perhaps because subgroups within the data are not sufficiently distinct from each other. Therefore, the results of the cluster analysis should be interpreted with a degree of caution.

Both clusters were fairly large, with one containing 69.4% of the cases ($n = 281$) and the other containing 30.6% of the cases ($n = 124$). An examination of the cluster centroids revealed that the difference between the two clusters was purely quantitative, with the larger cluster having lower scores than the smaller cluster on all four variables. Knowing that some students score more highly than others adds little to our understanding of adaptive number knowledge. At best, this solution would provide a justifiable division into "high" and "low" groups in the event that one would wish to do inter-group comparisons, which would be preferable to an arbitrary cut-off point (Hickendorff et al., 2018). However, that is not required in the present research. It was therefore decided to investigate whether a more meaningful solution could be found.

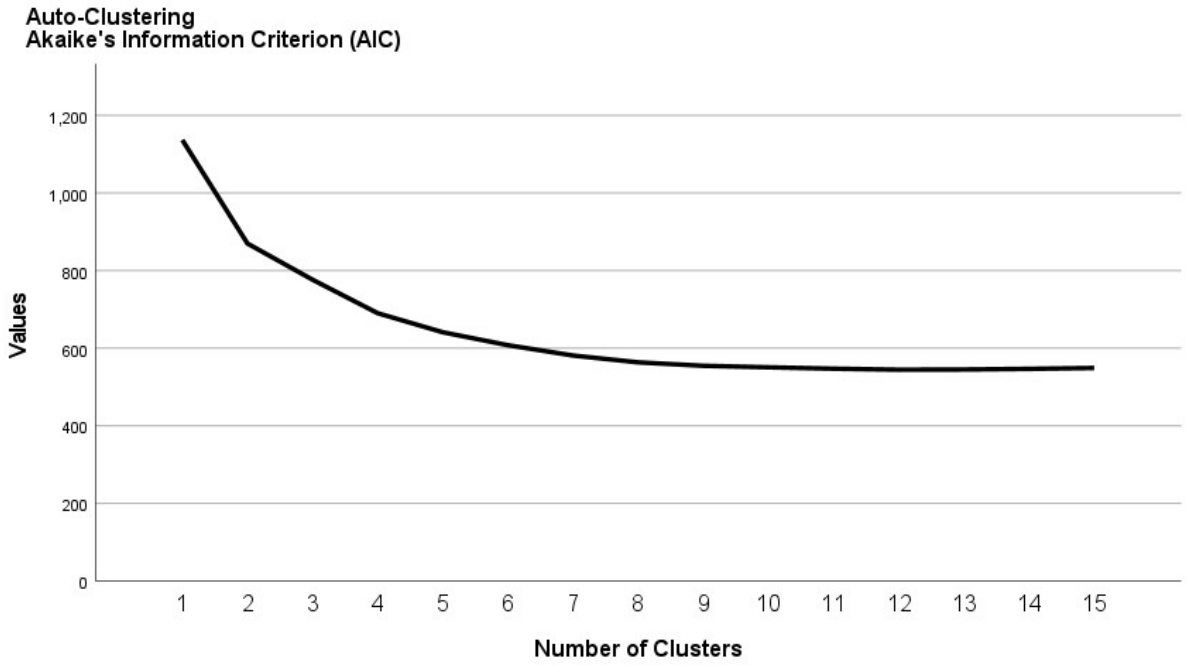
AIC measures the relative quality of a set of models; the preferred model is typically found when the AIC is minimised (Gaskin, 2015; Zajic, 2019). An inspection of the AIC values demonstrated that the minimum AIC value was at 12 clusters; this is obviously too large for an easily interpretable model. However, the graph of the AIC values at different cluster sizes (Figure 4.2) shows that there is a sharp improvement between the 1-cluster and 2-cluster solution, after which there is a visible "elbow" in the graph. This is probably why the 2-cluster solution was selected by SPSS. However, the AIC continues to decrease quite rapidly until the number of clusters is 4, where there is another "elbow", slightly less noticeable than the first. Thereafter, the AIC decreases at an increasingly gradual rate until it reaches its minimum at

[22] The silhouette measure of cluster cohesion and separation ranges from -1 to 1. A silhouette coefficient between -1 and 0.25 is classified by SPSS as "Poor", a coefficient between 0.25 and 0.5 is classified as "Fair", and a coefficient above 0.5 is classified as "Good". These cut-off points are based on the guidelines of Kaufman and Rousseeuw (IBM Corporation, 2021c).

the 12-cluster solution. This suggests that the 3- and 4-cluster models are also likely to add significantly to the explanatory power of the model, and may be worth considering.

Figure 4.2

Graph Showing Akaike's Information Criterion Against the Number of Clusters in the TwoStep Cluster Analysis model

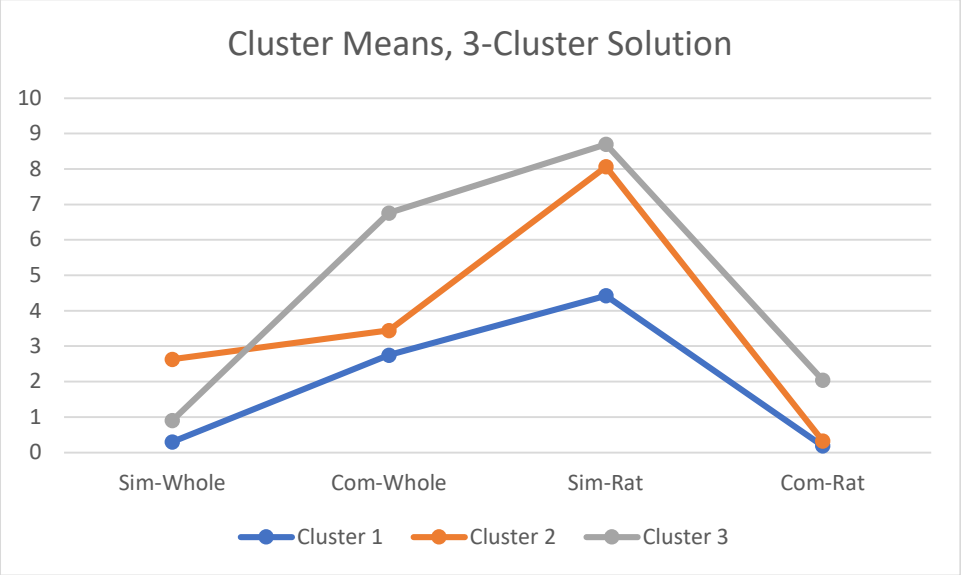


A second TwoStep cluster analysis was run on the four indicator variables. The procedure was the same as before, except that a three-cluster solution was specified. The fit of the three-cluster solution was also deemed to be "fair", with a silhouette coefficient of 0.4. This solution generated one large cluster (Cluster 1, 64.7% of cases, $n = 262$) and two much smaller ones (Cluster 2, 17.0% of cases, $n = 69$, and Cluster 3, 18.3% of cases, $n = 74$). This made the ratio of the largest cluster to the smallest cluster 3.80, which is above the level of 3.0 which Gaskin (2012) recommends as a rule of thumb. However, in the three-cluster model, new distinctions began to emerge between the groups. Cluster 1 still had the lowest scores on all indicator variables. Cluster 2 had the middle scores on most indicator variables, but the top score for the number of simple solutions generated on whole-number items. In fact, a closer examination (see Figure 4.3) revealed that while Cluster 1 scored low on all variables and Cluster 3 scored high on all variables, Cluster 2 scored low on complex solutions but high on simple solutions. In other words, this group looked very similar to the "Simple" profile

identified in McMullen et al.'s (2017, 2019) research with elementary school and lower secondary school students.

Figure 4.3

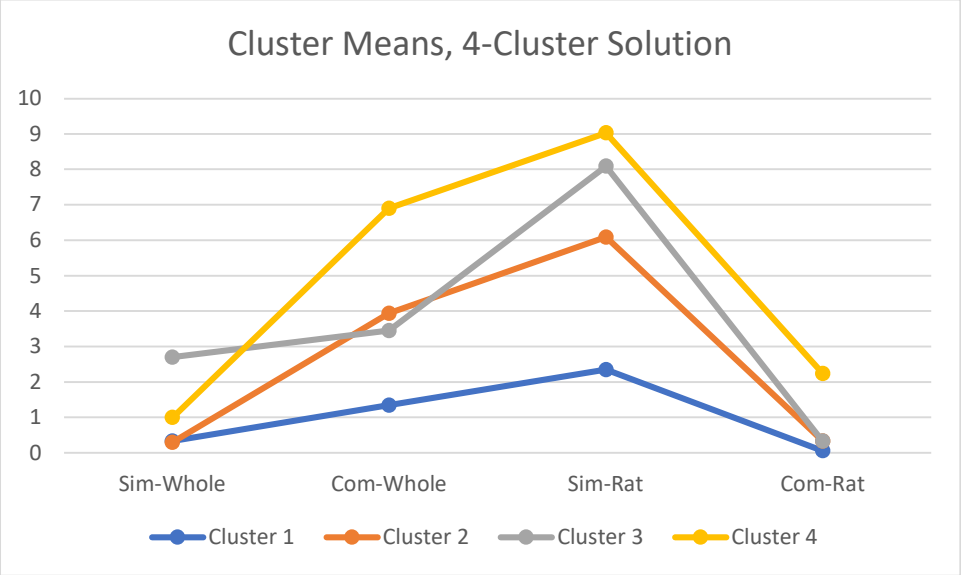
Graph Showing the Cluster Means for the Three-Cluster Solution



A third TwoStep cluster analysis was run on the four indicator variables, this time with a four-cluster solution specified. The fit of the four-cluster solution was also "fair", with a silhouette coefficient of 0.4. The division of cases between clusters was more even in this model: Cluster 1 (27.9%, $n = 113$), Cluster 2 (40.2%, $n = 163$), Cluster 3 (16.3%, $n = 66$), Cluster 4 (15.6%, $n = 63$). This resulted in a ratio of 2.59 between the largest cluster and the smallest cluster. Clusters 3 and 4 are very similar in size and pattern to the "simple" and "high" clusters identified in the previous model, while Cluster 1 has low scores across the board and Cluster 2 has mid-level scores across the board (see Figure 4.4). This is interesting in its own right, because in previous research of this type (McMullen et al., 2016, 2017, 2019), the low-scoring group has always been the largest. Even when McMullen et al. tested models with more groups, the result was typically a subdivision of one of the small, high-scoring groups, rather than a subdivision of the large, low-scoring group, which suggests that the low-scoring group is relatively stable and homogeneous. However, in this model, it is the middle group which is the largest. This suggests a somewhat different structure to the data collected in earlier research, possible reasons for which will be explored in Section 4.3.3.

Figure 4.4

Graph Showing the Cluster Means for the Four-Cluster Solution



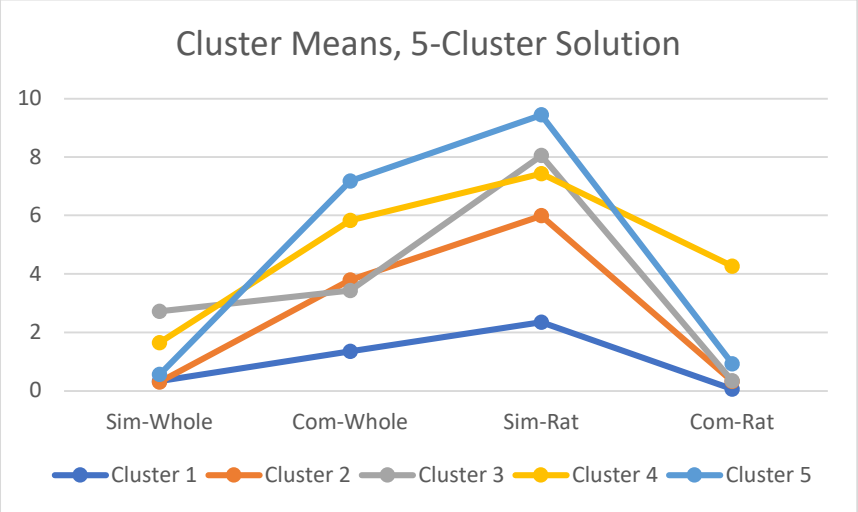
It seemed unlikely that a five-cluster model would improve interpretability much, as so far only one group (the "simple" group) was qualitatively different to the others; the other three clusters only differed quantitatively. However, for the sake of completeness, a five-cluster model was inspected to check that it was no better than the three- or four-cluster model.

A final TwoStep cluster analysis was run on the four indicator variables, this time with a five-cluster solution specified. The fit was once again deemed "fair", with a silhouette coefficient of 0.4. This model included one extremely small group, Cluster 4 (5.7% of cases, $n = 23$), which appeared to have resulted from a subdivision of the "high" group from the four-level model. The other clusters were as follows: Cluster 1 (27.9%, $n = 113$), Cluster 2 (37.5%, $n = 152$), Cluster 3 (16.0%, $n = 65$), Cluster 5 (12.8%, $n = 52$). Surprisingly, the very small Cluster 4 had interpretive value: the Cluster 4 scores were relatively high across all variables, and in particular members of this group produced many more complex solutions for the rational number items than members of any other group (Figure 4.5). However, the small size of this group also resulted in a ratio of 6.61 between the largest cluster and the smallest cluster.

Subsequent TwoStep cluster analyses with more clusters did not add any explanatory power; additional clusters were very similar to existing clusters. Therefore, the decision of which cluster model was the best came down to the three-, four- and five-cluster models.

Figure 4.5

Graph Showing the Cluster Means for the Five-Cluster Solution



The four-cluster model was deemed to have more interpretive value than the three-cluster model, because of its unusually small "low" group, as noted above. The four-cluster model also had a lower ratio between the largest and smallest group sizes. Nevertheless, small groups can be valuable to include in a cluster model, as they "may represent important groups of individuals with extreme [characteristics] that warrant attention" (Clatworthy et al., 2005, p. 353). This meant that it was not necessary to reject the five-cluster model simply because one of the groups was very small. The decision could be made, instead, on the basis of interpretability. The existence of the small group in the five-cluster model was interesting because of its unusually high scores on the Complex-Rational variable, a rare occurrence in the sample at large. It could therefore generate insights into the type of student who generates that type of solution. Therefore, it was decided to retain the five-cluster model for further analysis. Cluster membership was saved for each case in the data set.

4.3.3 Discussion of the Five-Cluster Model

The mean values for each variable in each cluster are given in Table 4.8. For ease of reference, the following names will henceforth be used to describe the five clusters: Cluster 1 – *Low*, Cluster 2 – *Medium*, Cluster 3 – *Simple*, Cluster 4 – *Complex*, Cluster 5 – *Strategic*. The Low group scores produced very few solutions of any type. The Medium group produced a roughly average number of solutions in each category. The Simple group produced an average or below-average amount of complex solutions, but a very high number of simple solutions

relative to the other clusters. The Complex group produces an above-average number of solutions in every category, and this is particularly noticeable on complex rational solutions: the mean for the Complex group in this category is 4.26 solutions, while for all other groups the mean figure is less than 1.[23] The Complex group generates the most solutions out of any group. Finally, the Strategic group is so-named because it seems to select the most efficient way to generate solutions for each question type: a very high number of complex solutions on the (sparse) whole number items, and a very high number of simple solutions on the (dense) rational number items, and relatively few other solutions.

Table 4.8

Mean Values for the Five-Cluster Model

	Simple Whole	**Complex Whole**	**Simple Rational**	**Complex Rational**	**Total**
Low (27.9%)	0.33	1.35	2.35	0.06	**4.09**
Medium (37.5%)	0.30	3.79	5.99	0.32	**10.40**
Simple (16.0%)	2.72	3.43	8.05	0.34	**14.54**
Complex (5.7%)	1.65	5.83	7.43	4.26	**19.17**
Strategic (12.8%)	0.56	7.17	9.44	0.92	**18.09**
Entire Sample (100%)	*0.81*	*3.60*	*5.83*	*0.55*	***10.79***

The differences between groups are partly in the absolute number of answers generated, but also in the ratio of simple to complex answers. Therefore, these ratios are given in Table 4.9 for whole number and rational number items. Table 4.9 makes it particularly clear that the Complex cluster produces a far higher proportion of complex solutions on the rational number items than any other group, although it still produces more simple solutions than complex solutions for these items. It also shows the reversal of the Strategic group's solution strategy when answering sparse whole items as compared to denser rational items.

In some ways, this model is similar to the profile and cluster models that have been found in previous research. Firstly, qualitative (as opposed to quantitative) differences in scoring patterns are found primarily among the higher-scoring participants. Previous research is unclear on why this should be the case. That said, it is noteworthy that in both the large-scale

[23] As a point of interest, only 25 students in the entire sample generated more than two complex solutions on the rational number items. Almost all of them must, logically, be in the 23-person Complex group.

studies which have performed a similar analysis of simple and complex solutions (McMullen et al., 2017, 2019) and in this study, the performance of the lower-scoring groups follows the same general pattern as the performance of the high-scoring Strategic group (just with lower absolute scores): more simple solutions on dense items, and more complex solutions on sparse items. This suggests that perhaps the default approach is to adopt a strategy which is adaptive to task characteristics, and only to vary it if one has developed particular skills which would make a different strategy more profitable. For instance, the Simple group clearly has a high facility with uni-operational calculations, and therefore they deploy this ability even on the sparse whole number items in order to generate as many solutions as possible.

Table 4.9

Ratios Between Simple and Complex Solutions for the Five-Cluster Model

*(Note that **different ratios** are presented for whole and rational items.)*

	Ratio of complex to simple solutions (whole)	**Ratio of simple to complex solutions (rational)**
Low (27.9%)	4.09	39.17
Medium (37.5%)	12.63	18.72
Simple (16.0%)	1.26	23.68
Complex (5.7%)	3.53	1.74
Strategic (12.8%)	12.80	10.26
Entire Sample (100%)	*4.44*	*10.6*

A second similarity to prior research is in the characteristics of certain groups. The Low, Simple and Strategic groups closely resemble profiles found in McMullen et al.'s (2017, 2019) prior research on ANK in Finnish elementary and lower high school students. This suggests that these three performance profiles may be consistent across countries and ages, although the proportion of the total sample that they make up may vary. The Complex group resembles McMullen et al.'s "High" profile (which is likewise very small size), producing an above-average number of solutions on all items, and the highest total score of any group.

There are also some differences to the profiles found in prior research. Primarily, as noted above, the Low group is unusually small. In the two large-scale studies mentioned in the previous paragraph, McMullen et al. found that the lowest-scoring group (called "Basic" in their publications) made up 50-60% of the total sample. In this case, the Low group only makes

up 27.9% of the sample, and there is a separate Medium group which makes up a further 37.5% of the sample, which has been completely absent from previous studies. The Medium group seems to be very clearly distinct from the Low group: the mean total score difference between the two groups is 6.31, which is larger than the difference between any two other consecutive group totals. In addition, the split between the Low and Medium group happened at a relatively early stage (in the four-cluster model), before the Complex group split from the highest-scoring group.

The scores in the Low group are remarkably low for students aged 14-18 years old, whom one would expect to have a reasonably good grasp of arithmetic. This group averaged just 1 correct answer per item. Looking specifically at the whole number items, which were identical to the sparse items used in McMullen et al. (2019), the Low group in the present study generated on average 1.68 correct solutions, while the equivalent group (Basic) in the 2019 study generated an average of 3.5 correct solutions. These groups are not directly comparable, since the Basic group in the 2019 study consisted of 58.8% of the total sample, instead of just 27.9% in the present study, and thus presumably included higher-scoring students as well as very low-scoring ones. It is nonetheless interesting to observe that the seventh-graders of the 2019 study significantly outperformed the much older students in the present study.[24] The reason for this poor performance cannot be determined with the present data, but possible explanations may include (1) a gap between high-achieving and low-achieving students that widens over time, causing the low-achieving students to stagnate and possibly even regress; (2) high math anxiety, potentially related to low achievement and low mathematical self-efficacy; or (3) simple disinterest and lack of effort in a not-for-marks test (particularly likely in this older age group, which is less likely to make an effort simply to please an adult, and in the test-centric academic culture which is anecdotally more prevalent in the USA than in Finland).

4.3.4 Associates of Cluster Membership

Based on the results of Section 4.2, it was expected that gender and mathematics module enrolment would be significantly related to cluster membership, and that age would not be.

[24] In fact, the seventh-grade sample of McMullen et al. (2019) slightly outperforms the present sample overall on these shared items. The seventh-graders produced an average of 4.98 solutions across these two items (1.90 simple and 3.08 complex), while the high schoolers in the present sample produced an average of 4.41 solutions across the same two items (0.81 simple and 3.60 complex).

School grade was expected to only be significantly associated with cluster membership for eleventh-graders. Chi-square tests were conducted to test these hypotheses. It was also expected that higher-scoring clusters would generate more cross-notational solutions on rational number items. A one-way ANOVA was conducted to test this hypothesis.

The first chi-square test showed a significant association between cluster membership and gender, $\chi^2(4,397) = 12.27, p = 0.02$. The Low cluster was disproportionately female (61.5% female), while the Complex and Strategic clusters were disproportionately male (Complex: 69.6% male, Strategic: 59.6% male).[25] No obvious gender disparity was apparent in the Medium or Simple clusters.

The second chi-square test showed a highly significant association between cluster membership and mathematics module enrolment, $\chi^2(20,405) = 111.03, p < 0.001$. Students in the Algebra 1 Honors, Geometry and Algebra 2 modules were disproportionately likely to be in the Low cluster, while students taking AP Statistics module were disproportionately likely to be in the Complex or Strategic clusters. Students in the Geometry Honors module were underrepresented in the Low cluster, but overrepresented in the Medium cluster. Students taking Algebra 2 Honors module seemed to be distributed between the clusters approximately as would have been expected at random, with only a slight tendency to be in higher-scoring clusters. The full crosstabulation of the chi-square analysis is given in Table 4.10. Note that, although none of the expected values are below 1, eight of the expected counts (27%) are below 5. This contravenes the recommendation that in a contingency table larger than 2 x 2, a maximum of 20% of expected counts should be below 5 (Field, 2018). If this requirement is not met, the power of the chi-square test decreases drastically (Field, 2018), which means that the test is more likely to make a Type II error, failing to reject the null hypothesis even if there is a real effect. In this case, a statistically significant effect has been found *even though* the test has reduced power, which suggests that the association between cluster membership and mathematics module enrolment is very strong indeed.

[25] Out of the 397 students who had valid data on all four indicator variables used in the cluster analyses *and* did not have missing gender data, 48.6% were female and 51.4% were male.

Table 4.10

Chi-Square Cross-Tabulation of Cluster Membership by Mathematics Module

		Alg. 1 Honors	Alg. 2	Alg. 2 Honors	AP Statistics	Geom.	Geom. Honors	ALL
Low Cluster	Count	27	23	30	0	12	21	**113**
	Expected count	11.7	14.0	33.5	8.1	5.0	40.7	**113**
	% within cluster	24%	20%	26%	0%	11%	19%	**100%**
	% within math module	64%	46%	25%	0%	67%	14%	**28%**
Medium Cluster	Count	12	15	42	8	4	71	**152**
	Expected count	15.8	18.8	45.0	10.9	6.8	54.8	**152**
	% within cluster	8%	10%	27%	5%	3%	47%	**100%**
	% within math module	29%	30%	35%	27%	22%	49%	**37%**
Simple Cluster	Count	0	10	20	4	2	29	**65**
	Expected count	6.7	8.0	19.3	4.7	2.9	23.4	**65**
	% within cluster	0%	15%	31%	6%	3%	45%	**100%**
	% within math module	0%	20%	17%	14%	11%	20%	**16%**
Complex Cluster	Count	0	1	9	6	0	7	**23**
	Expected count	2.4	2.8	6.8	1.6	1.0	8.3	**23**
	% within cluster	0%	4%	39%	26%	0%	31%	**100%**
	% within math module	0%	2%	7%	21%	0%	5%	**6%**
Strategic Cluster	Count	3	1	19	11	0	18	**52**
	Expected count	5.4	6.4	15.4	3.7	2.3	18.7	**52**
	% within cluster	6%	2%	36%	21%	0%	35%	**100%**
	% within math module	7%	2%	16%	38%	0%	12%	**13%**
Total	**Count**	**42**	**50**	**120**	**29**	**18**	**146**	**405**
	% of total	**11%**	**12%**	**30%**	**7%**	**4%**	**36%**	**100%**
	% within math module	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

The third chi-square test showed that there was no significant association between cluster membership and age, $\chi^2(16{,}393) = 19.89\,, p = 0.23$. The single 13-year-old was removed from the sample for this test, but eight cells (32%) still had expected counts smaller than 5, with one expected count smaller than 1. As noted above, this scenario results in reduced statistical power and possible Type II errors. It was impossible to run Fisher's exact test due to an insufficiently powerful computer, so instead a check was conducted by also removing the small 18-year-old group ($n = 13$) from the sample used to calculate the chi-square statistic. In this case, there were only two cells (10%) with expected counts under 5, and no cell had an expected count under 1. This test also returned a non-significant result, $\chi^2(12{,}381) = 14.06\,, p = 0.30$. This provides further support for the conclusion that there is no significant association between cluster membership and age.

The fourth chi-square test showed a significant association between cluster membership and school grade, $\chi^2(12{,}401) = 32.75\,, p = 0.001$. Using the rule that a standardised residual with an absolute value greater than 2 indicates a significant difference to the expected value (Field, 2018), it was found that eleventh-graders were disproportionately likely to be in the Low cluster ($z = 3.8$). No other cells had a standardised residual above 2. This is consistent with the findings in Section 4.2.2 that eleventh-graders performed more poorly than other students.

Before conducting the one-way ANOVA, the distribution of the number of cross-notational solutions was examined in the small Complex cluster ($n = 23$). The distribution showed no significant deviations from normality, and as the other clusters were all large enough (above 50) that the Central Limit Theorem was likely to apply, so the data were considered suitable for an ANOVA. The one-way ANOVA revealed a significant effect of cluster membership on the number of cross-notational solutions produced, $F(4{,}405) = 30.75$, $p < 0.001$. Levene's test indicated that the variances differed significantly between clusters, $F(4{,}400) = 17.88$, $p < 0.001$, so the Games-Howell post hoc procedure was used. The post hoc comparisons indicated that the Low cluster generated significantly fewer cross-notational solutions than any other cluster ($p < 0.001$), and the Complex and Strategic clusters additionally generated significantly more cross-notational solutions than the Medium cluster ($p < 0.05$). No other

significant differences were detected. The mean number of cross-notational solutions for each cluster is given in Table 4.11.

Table 4.11

Mean Number of Cross-Notational Solutions for Each Cluster

Cluster	Low	Medium	Simple	Complex	Strategic
Mean score	0.23	1.20	1.48	2.74	2.04

The results of the chi-square tests therefore confirmed the hypotheses that gender and mathematics module enrolment were significantly related to cluster membership, that age was not significantly related to cluster membership, and that school grade was only significantly associated with cluster membership for eleventh-graders. The one-way ANOVA confirmed the hypothesis that high numbers of cross-notational solutions were associated with membership of the highest-scoring clusters (Complex and Strategic), and that lower numbers of cross-notational solutions were associated with membership of lower-scoring clusters (Low and Medium). Because these results are consistent with the findings in Sections 4.1 and 4.2, they also provide external validation for the five-cluster model (Aldenderfer & Blashfield, 1984).

4.4 Conclusion

This chapter has generated a number of interesting findings about the individual differences in adaptive whole and rational number knowledge in the present sample. These findings are related to the development of A(R)NK over time, the interaction of gender and A(R)NK, the underlying five-cluster structure of the sample, and the use of cross-notation on rational number items. In addition to these substantive findings, some methodological observations have been made which may be useful for future research.

Firstly, the analyses of inter-group differences in Section 4.2 contribute to a better understanding of how A(R)NK develops over time. It seems that achievement on the ASP task is not related to age in this sample, despite the fact that studies with younger children have shown a strong association between school grade and number of solutions produced. This may suggest that A(R)NK growth levels off in the teenage years, or alternatively that A(R)NK patterns differ between Finland and the USA. If the former proves to be true, this would be a

significant addition to the collective understanding of A(R)NK development. It also appears that performance on the ASP task is strongly related to mathematics module enrolment, with students in more advanced courses performing better on the ASP task as well. This suggests that the absence of age-related differences in A(R)NK could also be due to the fact that, in this sample, age is not tightly bound to the level of mathematical instruction and exposure a student has had (unlike previous samples from Finnish elementary and lower secondary schools). It is not, however, clear whether participation in more advanced mathematics courses fosters A(R)NK or vice versa, or if both can be explained by some third variable. Future longitudinal studies are therefore recommended.

A clear link between school grade and achievement on the ASP task was not found, giving support to the possibility that it is mathematical experience, rather than age, which has a stronger bearing on A(R)NK levels. However, it was found that the eleventh-grade subsample performed significantly worse than students from any other grade. This could be related to the fact that the eleventh-graders in this sample have lower enrolment rates in Honors and AP courses, and draws attention to the fact that most of the present sample *is* enrolled in Honors and AP courses. This means that the present research should be generalised with caution, and is likely to be most applicable to groups of students who are taking relatively advanced mathematics courses for their grade.

Second, the inter-group comparisons revealed an unexpected association between gender and performance on the ASP task, with boys significantly outperforming girls. This is a completely novel finding which might open the door to a deeper understanding of differential development of A(R)NK and mathematical expertise more broadly in males and females. One possibility is that boys do generally have higher levels of arithmetic adaptivity than girls. If this is true, multiple explanations are possible. One would be that male brains have some greater capacity for adaptive mathematical thinking than female brains, in much the same way that they seem to be better at spatial reasoning tasks (Voyer et al., 1995, 2000). A sociocultural explanation would also be possible. It is widely speculated that the modern schooling system does not suit boys well, as countless articles lamenting “the boy problem” suggest. However, a side-effect of this state of affairs may be that boys are more “robust to schooling”, in that they are less likely to internalise the routinised approach to mathematics expected by teachers and standardised tests, and more likely to preserve the more creative approach to mathematical

problem-solving we often see in young children.[26] A different possibility is that boys do not in fact have higher levels of arithmetic adaptivity than girls, but task features have resulted in their scoring higher on the present test. One possible explanation along these lines is that the ASP task items were all performed under time pressure and test conditions. Girls tend to have higher levels of mathematics anxiety (Dowker et al., 2016), which could affect their performance negatively in a time-pressured test situation, so it is possible that this test format resulted in girls scoring lower than they would otherwise have done. All of these possible explanations are mere speculation at this point; further research is needed to establish which, if any, explains the gender gap in performance on this ASP task.

Third, the five-cluster model selected in Section 4.3 demonstrated a latent group structure that bears a substantial resemblance to previous research. In particular, qualitative differences in answer patterns only emerged among higher-scoring students: specifically, students in the Simple and Complex clusters demonstrated a higher propensity to use the solution types their clusters are named for. This paper posits the suggestion that the default approach for all students seems to be to adopt a strategy which is adaptive to task characteristics (using complex solutions for sparse items and simple solutions for denser items), and that this approach only varies if a student has developed particular expertise in a particular solution strategy. This suggestion is supported by all large-scale studies to date, but as there have only been three large-scale studies including this one, it can be regarded as tentative at best. The five-cluster model also revealed a clear split between low-achievers and middle-achievers which has not been found in research on younger, Finnish students. The cause of the extreme low scores in the Low cluster could potentially be due to a growing achievement gap, or alternatively to a lack of interest in the testing process. The validity of this five-cluster model is bolstered by the fact that membership of high-scoring profiles is significantly associated with mathematics module enrolment, gender, and the number of cross-notational solutions (see below).

A fourth substantive finding came from the initial examination of descriptive statistics: barely half the sample generated even a single solution which combined fractions and decimals. This is likely to indicate an incomplete understanding of the connection between these two representations of rational numbers, which would imply gaps in the network of knowledge that

[26] Thanks must go to my supervisor, Koen Veermans, for the phrase “robust to schooling”.

underpins high levels of A(R)NK. This interpretation is strengthened by the fact that use of cross-notational solutions is strongly associated with membership of more adaptive clusters.

In addition to these substantive findings, two of the observations made in this chapter may be valuable from a methodological perspective. The first observation is that the maximum number of solutions in the ASP task increased for each subsequent item, although the mean score did not. If this is a consistent trend in all research using the ASP task, this should be taken into consideration when deciding on the order of items (e.g. by alternating item types) and when interpreting the results of high-achievers. It is therefore recommended that other data collected using the ASP task is examined for evidence of such a trend. However, since this effect seems to be confined to top achievers and does not significantly affect the mean scores, there seems little cause to be concerned about testing effects increasing the overall results on later items. The second observation is that substantially more solutions were produced for the rational number items than the whole number items. As this is an improbable result, given what is known about students generally finding rational arithmetic more difficult than whole number arithmetic, it speaks to the importance of including comparable whole number and rational number items in future tests. In this test, the whole number items were sparse, using relatively uncommon target numbers with few one-step solutions, while the rational number items were denser, with many obvious one-step solutions. This likely enabled the production of a disproportionately high number of responses on the rational number items. Including a balance of dense and sparse whole and rational items in future tests would allow a better comparison of adaptive expertise with whole and rational numbers; it would also enable an assessment of whether the low number of complex responses on rational items was due to their denseness, or due to discomfort with more demanding rational arithmetic.

5. Evaluation of a Multi-Task Instrument for Measuring A(R)NK

This chapter addresses the third research question, namely: *Does an instrument consisting of three different tasks provide a more nuanced understanding of A(R)NK than the ASP task provides alone?* The three tasks under consideration are the ASP task, the Missing Symbol task and the Rapid Verification task.

The first two sections of this chapter consider descriptive statistics for the Missing Symbol and Rapid Verification tasks as a starting point for a general discussion of how these tasks performed. As the ASP task has already been considered extensively in previous chapters, it is not revisited in these sections. The third section uses an EFA to investigate whether all three tasks can be considered to measure a single latent construct. If the EFA were to identify a single latent variable, this would support the hypothesis that all three tasks measure adaptive number knowledge.

5.1 Discussion of Missing Symbol Task

5.1.1 Whole Number Items – Descriptive Statistics

Descriptive statistics for the whole number items in the Missing Symbol task are given in Table 5.1. It is immediately apparent that several of the whole items have extremely high average values. Items W1, W2 and W5 all have means over 0.9, which indicates that over 90% of participants answered these items correctly. It is important to interrogate why these particular items seem to have been very easy for the participants: firstly, because it may generate insights into the general level of adaptive number knowledge in students in this sample, and secondly, because a measurement instrument must elicit a reasonably large degree of variation in order to be able to discriminate usefully between participants on the construct of interest – thus, identifying *why* certain items are less good at doing this is useful when evaluating a new instrument. One possible explanation is that these three items featured two-step equations – that is, equations containing only two operators, like $11 \cdot 6 - 7 = 59$ (W1) – while most other whole number items contained three-step equations, like $11 \cdot 5 + 2 \cdot 2 = 59$ (W3). The only other whole number item which featured a two-step equation was W8, which was the next-highest scoring item. It was answered correctly by almost 82% of participants. This suggests

that two-step items in the Missing Symbol task were significantly easier for the participants than three-step items, possibly because they involved fewer relationships between numbers and operations. It is possible that using three-step equations in future editions of the Missing Symbol task would elicit more varied responses from the participants. A second possible explanation for the high scores on these items is that W1 and W5 required a single blank space to be filled in, while the other whole number items required two blank spaces to be filled in – again, a simpler task which required fewer possibilities to be tested mentally before settling on the correct solution. However, this explanation is less likely because it does not explain why W2 (which contained two blank spaces) was also answered correctly by over 95% of participants. It also does not account for W3, which contains only one blank space but was only answered correctly by 72% of participants.

Table 5.1

Descriptive Statistics for the Missing Symbol Task (Whole Number Items)

	N	Min	Max	Mean	Std. Deviation	Blank Spaces
Missing Symbol W1	447	0	1	0.96	0.19	Op
Missing Symbol W2	447	0	1	0.96	0.21	Op, Op
Missing Symbol W3	447	0	1	0.72	0.45	Num
Missing Symbol W4	447	0	1	0.34	0.48	Num, Op
Missing Symbol W5	447	0	1	0.91	0.29	Op
Missing Symbol W6	447	0	1	0.63	0.48	Op, Op
Missing Symbol W7	447	0	1	0.49	0.50	Num, Op
Missing Symbol W8	447	0	1	0.82	0.39	Op, Op

The two lowest-scoring items are W4 (answered correctly by 34% of participants) and W7 (answered correctly by 49% of participants). These two items required participants to fill in one number and one operation, suggesting that this combination of missing symbols was particularly challenging. When only numbers are missing, the relationship between the missing values is already given by the operations that appear in the test item, so the participant needs only to figure out *which numbers* bear that particular relation to each other. When only operators are missing, the numbers are given and the participant needs only to work out *what relationship* they bear to each other. However, when both an operator and a number are missing, the participant must do both simultaneously: a more challenging task, which requires

a more robust mental network of connections between numerical characteristics and arithmetic relations. This suggests that such items are likely to be capable of differentiating between higher levels of adaptive number knowledge.

5.1.2 Whole Number Items – Constructing a Subscale

As the individual items in the Missing Symbol task had a very limited range of possible outcomes (0 or 1), they were unlikely to be good indicators for an exploratory factor analysis. Sum scores of the Missing Symbol items would be more useful, as they would have a larger range. This would maximise variance, which in turn would increase the common variance available in the EFA. As noted in Chapter 3, a higher ratio of variables per factor tends to result in better results (assuming a constant sample size). Since the number of factors was unknown, it seemed wise to create sum scores with an eye to having as many variables as possible in the final EFA. This suggested that separate sum scores should be created for the whole and rational number Missing Symbol items.

It would in fact be possible to create separate sum scores for each set of Missing Symbol items with the same target number (59, 38, ½, 3). However, each of these sum scores would have a small range (0-3 for target number 3 and 0-4 for the other target numbers), so it was deemed preferable to use a single sum score for whole number items and a single sum score for rational number items. This decision was bolstered by an examination of the score distribution for the Missing Symbol items with target number 59 (Figure 5.1). Very few participants scored 0 or 1 on these four items, making the effective range even smaller at 2-4. When the scores for all the whole number items are combined, the distribution shows much more variation (Figure 5.2), which is likely to be more useful in an EFA.

Furthermore, since items W1 and W2 (and, to a lesser extent, W5) were answered correctly by almost all participants, it needed to be considered whether they provided any useful differentiation between participants. If not, they could be omitted from the sum score scale. An examination of the reliability statistics did not provide strong reason to omit them from the scale. Cronbach's alpha across all eight whole number items was 0.67. If W1 and W2 were deleted, Cronbach's alpha increased only slightly to 0.68. If W5 was also deleted, Cronbach's

Figure 5.1

Histogram Showing Score Distribution for Missing Symbol Items (Target 59)

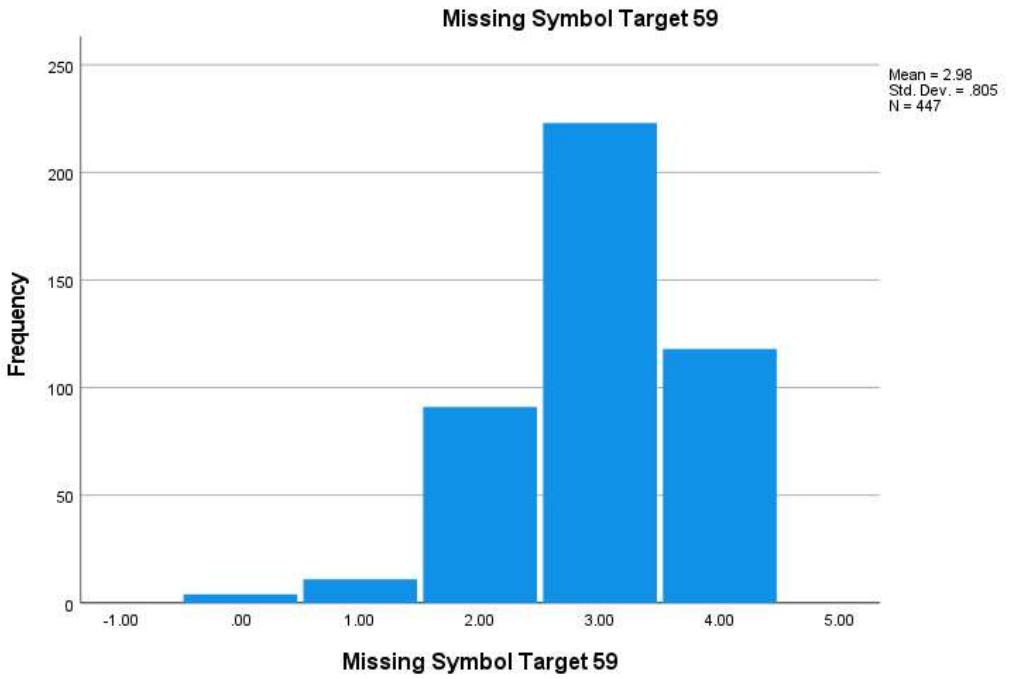
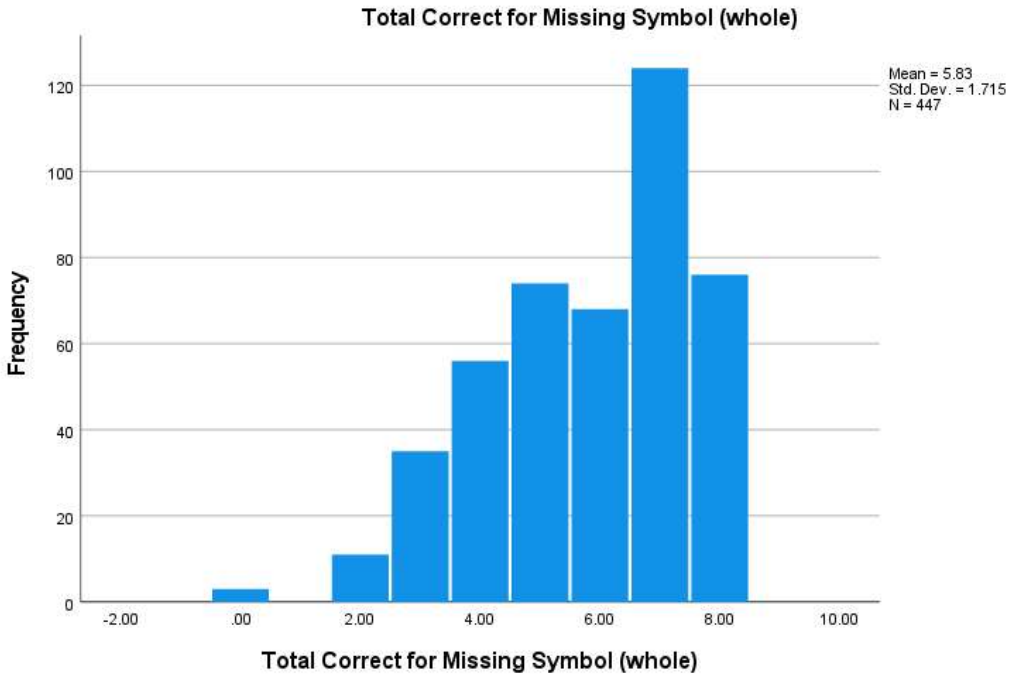


Figure 5.2

Histogram Showing Score Distribution for Missing Symbol Items (Targets 59 & 38)



alpha barely changed, still equalling 0.67 when corrected to two decimal places.[27] These tiny changes in the alpha value suggested that removing the items would not result in a significantly higher or lower degree of internal consistency in the whole number subscale. In other words, they were measuring the same thing as the other items; participants just found them very easy.

To gain further clarity, it was decided to compare the score distribution for all answers to the score distributions obtained when excluding the high-scoring items. Figure 5.3 presents the score distribution with items W1 and W2 removed. It can be seen that this histogram is virtually identical to the rightmost seven bars of Figure 5.2 (which is based on all eight whole number items). This confirms that items W1 and W2 do not add any discriminatory value; they merely inflate virtually all scores by two points. Figure 5.4 presents the score distribution with items W1, W2 and W5 removed. It is again very similar to the distributions in Figures 5.2 and 5.3, but it loses a little more discriminatory power at the lowest end of the scoring spectrum: whereas Figure 5.3 can distinguish between the 9 students who got no answers correct and the 30 who got one answer correct, Figure 5.4 lumps most of them together in the zero-score bracket. Even though A(R)NK is typically regarded as a characteristic of high achievers, it would still be desirable to retain high-resolution information at all scoring levels of the instrument, so that as accurate an understanding of the construct as possible can be obtained. It was therefore decided to exclude items W1 and W2 from the sum score for whole number items, but to include item W5. The descriptive statistics for the revised whole number subscale can be found in Table 5.2.

Table 5.2

Descriptive Statistics for the Revised Whole Number Subscale in the Missing Symbol Task

	N	Range	Min	Max	Mean	Std. Deviation
Missing Symbol revised whole subscale (excl. W1, W2)	447	6	0	6	3.91	1.62

[27] The figures to three decimal places were as follows: 0.668 for all eight items, 0.677 if W1 and W2 were excluded, 0.666 if W1, W2 and W5 were excluded.

Figure 5.3

Histogram Showing Score Distributions for Missing Symbol Items (Targets 59 & 38), excluding W1 and W2

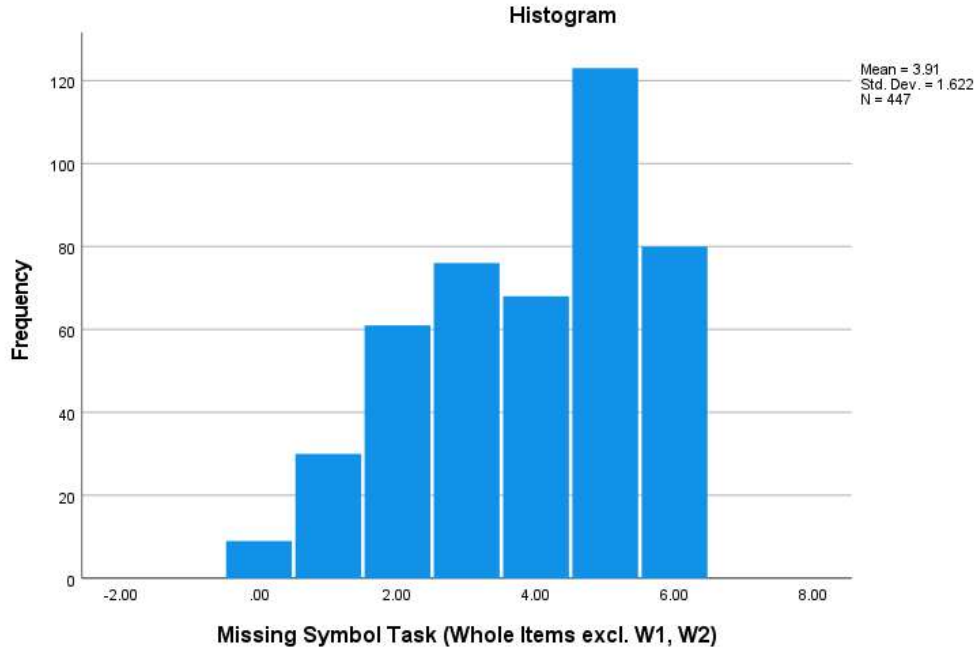
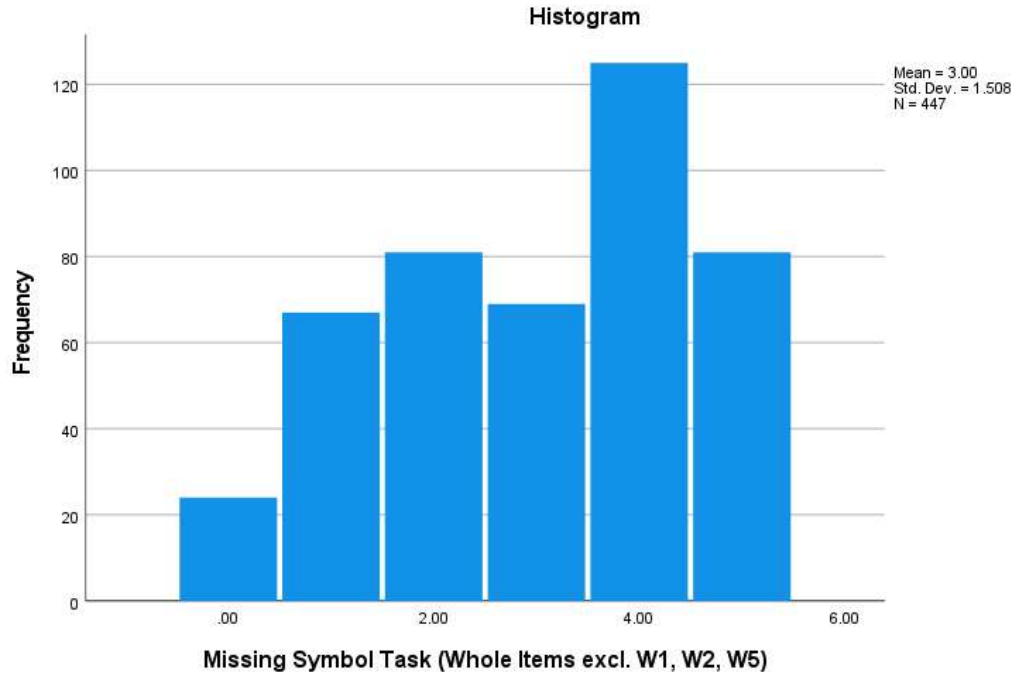


Figure 5.4

Histogram Showing Score Distributions for Missing Symbol Items (Targets 59 & 38), excluding W1, W2 and W5



5.1.3 Rational Number Items – Descriptive Statistics

Descriptive statistics for the rational number items in the Missing Symbol task are given in Table 5.3. In the Missing Symbol task, the rational number items clearly conformed to the expectation that participants would score lower on tasks involving rational numbers than those involving exclusively whole numbers. The scores across the board were noticeably lower than for the whole number items, with means ranging from 0.26 (Item R7) to 0.76 (Item R1). This was the case despite the fact that the rational number items were considerably less complex in form than the whole number items. Only one item out of the seven required participants to fill in two blank spaces; the others had only one blank space. By contrast, five out of the eight whole number items (and four out of the six items retained for the subscale) required participants to fill in two blank spaces. Similarly, four out of seven rational number items were one-step equations, like $0.75 - \frac{1}{4} = \frac{1}{2}$, and the remaining three items were two-step equations. By contrast, the whole number items contained no one-step equations, four two-step equations (of which only two were retained), and four three-step equations (which were all retained).

Table 5.3

Descriptive Statistics for the Missing Symbol Task (Rational Number Items)

	N	Minimum	Maximum	Mean	Std. Deviation	Blank Spaces
Missing Symbol R1	447	0	1	0.76	0.43	Op
Missing Symbol R2	447	0	1	0.33	0.47	Op
Missing Symbol R3	447	0	1	0.37	0.48	Op
Missing Symbol R4	447	0	1	0.46	0.50	Op, Op
Missing Symbol R5	447	0	1	0.66	0.48	Op
Missing Symbol R6	447	0	1	0.30	0.46	Op
Missing Symbol R7	447	0	1	0.26	0.44	Num

The finding that participants scored lower on rational number items is consistent with previous scholarship which has found that students find rational numbers more difficult to understand and calculate with than whole numbers (e.g. Lortie-Forgues et al., 2015; Siegler et al., 2011). This may suggest that this task succeeded better in setting comparable questions across number types than the ASP task, in which participants scored significantly higher on rational number items than on whole number items (see Section 4.1.1). It is likely also reflective of the fact that in the ASP task, participants were free to choose their own calculation strategy, which meant that they could avoid numerical combinations and operations that they were not comfortable

with. The Missing Symbol task was more constrained and thus offered less opportunity to avoid challenging numbers and procedures.

One piece of evidence in support of this second hypothesis was the fact that the four lowest-scoring rational number items (R2, R3, R6 and R7) – which fewer than 40% of participants answered correctly – were the ones which included or required division by a fractional value. This suggests that dividing by a fraction (or using division in an equation that contains fractions) is something that the participants found particularly difficult, potentially because the fraction division algorithm bears so little resemblance to whole number division. A direct comparison to the ASP task cannot be achieved without a qualitative analysis of the frequency of different solutions (which would be an interesting topic for future research), but it seems likely that participants who were able to avoid fractional division in the ASP task found themselves forced to engage with it in the Missing Symbol task. This may go some way towards explaining why participants performed relatively better on the rational number items in the ASP task but not in the Missing Symbol task.

Incidentally, it should be noted that none of the whole number items included or required a division symbol. This is one of several formal differences between the whole number items and the rational number items: as mentioned above, the whole number items also had more blank spaces to fill in and involved longer (two-step or three-step) equations. Although this seems to have resulted in the expected outcome, which is that participants score more highly on whole number items than rational number items, it might be informative to test directly equivalent whole number and rational number items in future iterations of this instrument. This may help to ascertain the optimal balance of formal complexity between whole number and rational number items; it may also generate insights into which facets of A(R)NK are most strongly developed in students in a particular age range.

5.1.4 Rational Number Items – Constructing a Subscale

Cronbach's α across all seven rational number items was 0.66. Since the analysis did not indicate that deleting items would improve reliability, and since none of the items exhibited severely limited variability, it was decided to retain all seven items. The reliability statistic is a little low, but acceptable since the component items are all dichotomous, as discussed in Section 2.2.2. A sum score variable was computed for the seven items. Descriptive statistics

for the subscale can be found in Table 5.4. An examination of the score distribution for this subscale confirms that the values are approximately normally distributed across the entire range (Figure 5.5).

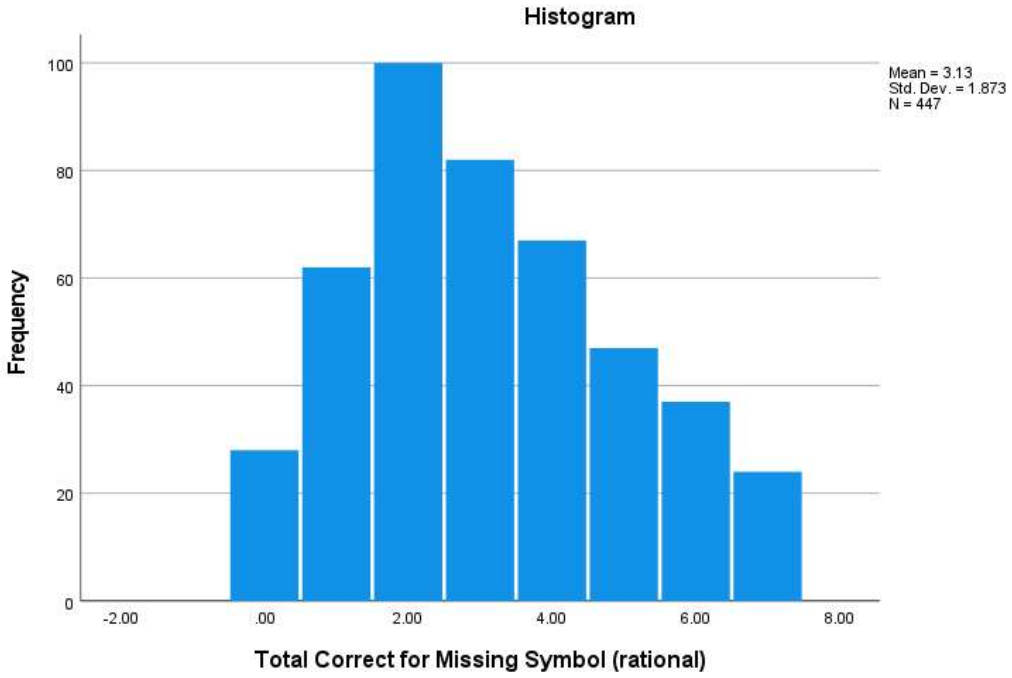
Table 5.4

Descriptive Statistics for the Rational Number Subscale in the Missing Symbol Task

	N	Range	Min	Max	Mean	Std. Deviation
Missing Symbol rational subscale	447	7	0	7	3.13	1.87

Figure 5.5

Histogram Showing Score Distributions for Missing Symbol Items (Targets ½ & 3)



5.2 Discussion of Rapid Verification Task

5.2.1 Whole Number Items – Descriptive Statistics

Descriptive statistics for the whole number items in the Rapid Verification task are given in Table 5.5. The True items (i.e. items which equalled the target number) are shaded in green and the False items (i.e. items which did not equal the target number) are shaded in yellow. It

is clear from the table that participants were far more likely to answer False items correctly: the mean for every False item is above 0.9; all but one False items have a mean of 0.95 or greater when rounded to two decimal places. This means that 95% of participants answered almost all of the False items correctly. By contrast, the means for the True items range from 0.27 to 0.84. They show much more variation than the False items, and the number of participants answering each one correctly is lower than for the True items.

A chi-square analysis was conducted on each trial (i.e. each set of items with a common target number) to quantify the difference in answer patterns between True and False items. Both chi-square analyses showed that False items were much more likely to be answered correctly than True items. For target number 59, $\chi^2(1,447) = 875.64, p = 1.9 \times 10^{-192}$. For target number 38, $\chi^2(1,447) = 1270.89, p = 2.4 \times 10^{-27}$. These chi-square statistics are so large, and the p-values correspondingly so tiny, that it can be seen there was an extremely strong relationship between whether an item was True or False, and whether a participant would answer it correctly.

This was a surprising finding, because there was no obvious difference in form or difficulty between the True items and the False items. There are several possible explanations, which are not necessarily mutually exclusive. One option is that participants may find it easier to determine whether an expression is *not* equal to a target number than to confirm that it *definitely is* equal to a target number. Rejecting an incorrect expression may require less precision: for instance, a general sense that the value would be much larger than the target number would be sufficient to reject a False item; recognising that the value would be even would also be sufficient to reject a False item if the target number were odd. By contrast, confirming that an expression is equal to the target number would require the participant to calculate an exact value for the expression. This might be more difficult to do altogether, or it might simply be more time-consuming than rejecting an incorrect answer, therefore resulting in lower correct rates in this time-limited task.

Table 5.5

Descriptive Statistics for the Rapid Verification Task (Whole Number Items)

	N	Mean	Std. Deviation	Skewness	Item Type
Rapid Verification Task 59a	447	0.95	0.21	-4.30	False
Rapid Verification Task 59b	447	0.90	0.30	-2.71	False
Rapid Verification Task 59c	447	0.78	0.42	-1.33	True
Rapid Verification Task 59d	447	0.95	0.22	-4.07	False
Rapid Verification Task 59e	447	0.84	0.36	-1.90	True
Rapid Verification Task 59f	447	0.97	0.17	-5.40	False
Rapid Verification Task 59g	447	0.63	0.48	-0.55	True
Rapid Verification Task 59h	447	0.97	0.18	-5.20	False
Rapid Verification Task 59i	447	0.97	0.17	-5.62	False
Rapid Verification Task 59j	447	0.95	0.23	-3.97	False
Rapid Verification Task 59k	447	0.97	0.17	-5.62	False
Rapid Verification Task 59l	447	0.34	0.48	0.67	True
Rapid Verification Task 38a	447	0.83	0.38	-1.74	True
Rapid Verification Task 38b	447	0.66	0.47	-0.69	True
Rapid Verification Task 38c	447	0.64	0.48	-0.57	True
Rapid Verification Task 38d	447	0.98	0.16	-6.16	False
Rapid Verification Task 38e	447	0.95	0.23	-3.97	False
Rapid Verification Task 38f	447	0.98	0.12	-7.83	False
Rapid Verification Task 38g	447	0.98	0.12	-7.83	False
Rapid Verification Task 38h	447	0.95	0.21	-4.30	False
Rapid Verification Task 38i	447	0.98	0.14	-6.86	False
Rapid Verification Task 38j	447	0.60	0.49	-0.40	True
Rapid Verification Task 38k	447	0.35	0.48	0.65	True
Rapid Verification Task 38l	447	0.27	0.44	1.04	True

A second option is that time pressure may have meant that fewer participants had time to consider the later items thoroughly. The final item in each trial was a True item, and so the low scores on these final items might signify that many participants simply ran out of time to answer them properly. However, this is unlikely to be the case, partly because True items earlier in the trials are also answered incorrectly more often than False items, and partly because False items very close to the end of each trial (such as 59k) were answered overwhelmingly correctly. If time pressure did mean that participants failed to consider every question, therefore, this had a

minimal impact on the differential True-False answer patterns. (It is, however, noticeable that the correct rates for True items tended to be higher on items that appeared earlier in each trial, so time constraints may have played a limited role.)

A third option is that participants had a non-circling bias; in other words, they adopted a strategy of not circling items unless there was a very compelling reason to do so. A high threshold for circling items could be consistent with the first possibility discussed above (that ascertaining whether a True item is equal to the target number may require more precision and time than rejecting a False item), as it might reduce the chance of making a mistake. In order to investigate whether a non-circling bias was present, the *response bias* was calculated for each participant. Response bias, represented by the letter *C*, is a key variable in signal detection theory (for more about SDT, see Section 2.2.3) and it represents a respondent's tendency to answer positively or negatively to dichotomous items (Geary et al., 2009; MacMillan, 2002). A *C*-value of zero is assigned to the "ideal observer", who adopts a strategy which minimises the chances of both a Miss and a False Alarm (Abdi, 2010).[28] A positive response bias indicates a high threshold for circling items; in other words, the respondent has a conservative strategy, circling items less often than the ideal observer. A negative response bias indicates a low threshold for circling items; such a respondent has a liberal strategy, circling items more often than the ideal observer. A conservative strategy would result in fewer False Alarms but more Misses, while a liberal strategy would result in more False Alarms but fewer Misses.

Using SDT to investigate participants' response biases was particularly advantageous since participants had not been required to write anything to indicate that they thought an item was incorrect. This meant that an uncircled item might indicate an active rejection, or alternatively it might indicate that a participant had simply not considered that item. If the *C*-values revealed a general non-circling tendency, this would support the possibility that participants were exercising a rejection bias rather than skipping items at random.

Response bias is calculated with the following formula: $C = -\frac{1}{2}(z_{Hits} + z_{FA})$. The variable z_{Hits} expresses the location of an individual's threshold for circling a True item, relative to a

[28] Recall from Section 2.2.3 the following definitions: A True item that was circled was coded as a Hit, a True item that was not circled was coded as a Miss, a False item that was circled was coded as a False Alarm, and a False item that was not circled was coded as a Correct Rejection.

theoretical distribution of the *signal intensity* (in this case, the signal would be a belief in the correctness of the item, generated by accurate mathematical assessments). The variable z_{FA} expresses the location of an individual's threshold for circling a False item, relative to a theoretical distribution of *noise* (in this case, the noise would be a belief in the correctness of the item, generated by random effects or other stimuli). As Abdi (2010) explains, stimuli generated by noise are assumed to be normally distributed with a mean of 0 and a standard deviation of 1. The signal is added to the noise, resulting in an identically-shaped signal distribution with a larger mean. The more easily the signal can be distinguished from the noise (i.e. the stronger the signal), the further apart the two curves will be.

If a participant has (say) a 0.8 probability of correctly circling True items, z_{Hits} is defined as the value on the normal distribution (with $\mu = 0$ and $\sigma = 1$) which has an associated probability of 0.8. Similarly, if a participant has a 0.3 probability of incorrectly circling False items, z_{FA} is defined as the value on the normal distribution which has an associated probability of 0.3. It is important to note that these z-scores are calculated relative to a theoretical normal distribution. They are not calculated relative to the distribution of responses in the sample.

Table 5.6

Descriptive Statistics for Response Bias (Rapid Verification Task, Whole Number Items)

	N	Min	Max	Mean	Std. Deviation	Skewness
Response bias (C) for Rapid Verification task 59	441	-1.91	4.50	1.21	1.30	-0.36
Response bias (C) for Rapid Verification task 38	440	-1.77	4.50	1.77	0.99	-0.95

Descriptive statistics for the response bias for each whole number trial are given in Table 5.6. The mean response bias was positive for both trials, indicating that the average threshold for circling items was 1.2 SD above the ideal observer's threshold for target 59, and 1.8 SD above the ideal observer's threshold for target 38. The histograms in Figures 5.6 and 5.7 illustrate that the positive results were not merely mathematical artefacts which obscured negative scores: very few individuals had a negative response bias. This suggests that the participants adopted a generally conservative strategy, that is, a rejection bias.

Figure 5.6

Histogram Showing Distribution of Response Bias for the Rapid Verification Task, Target 59

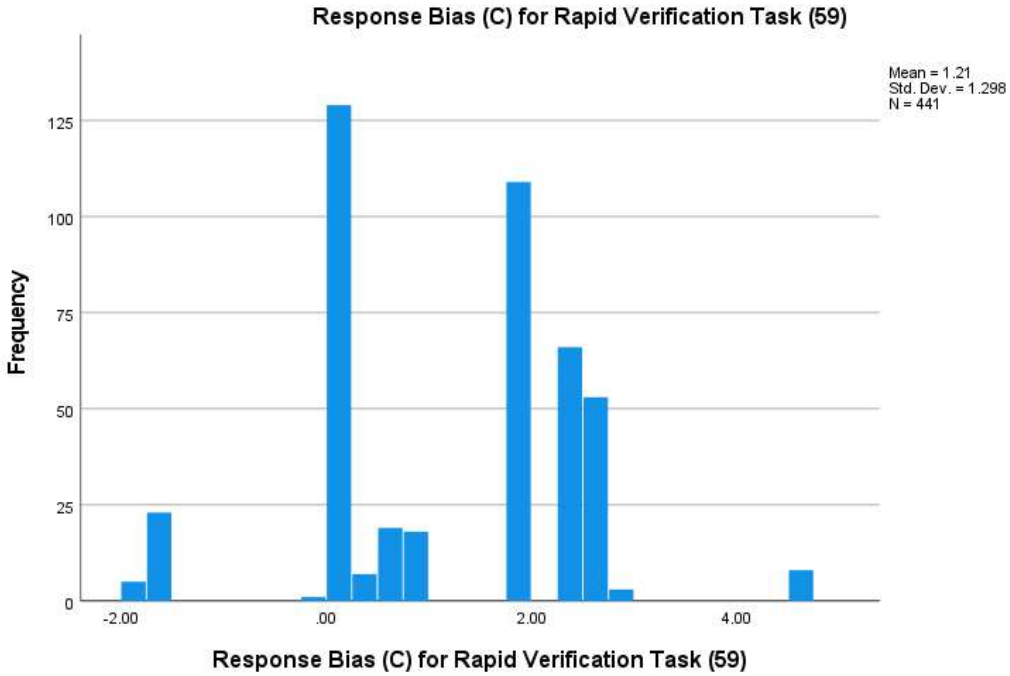
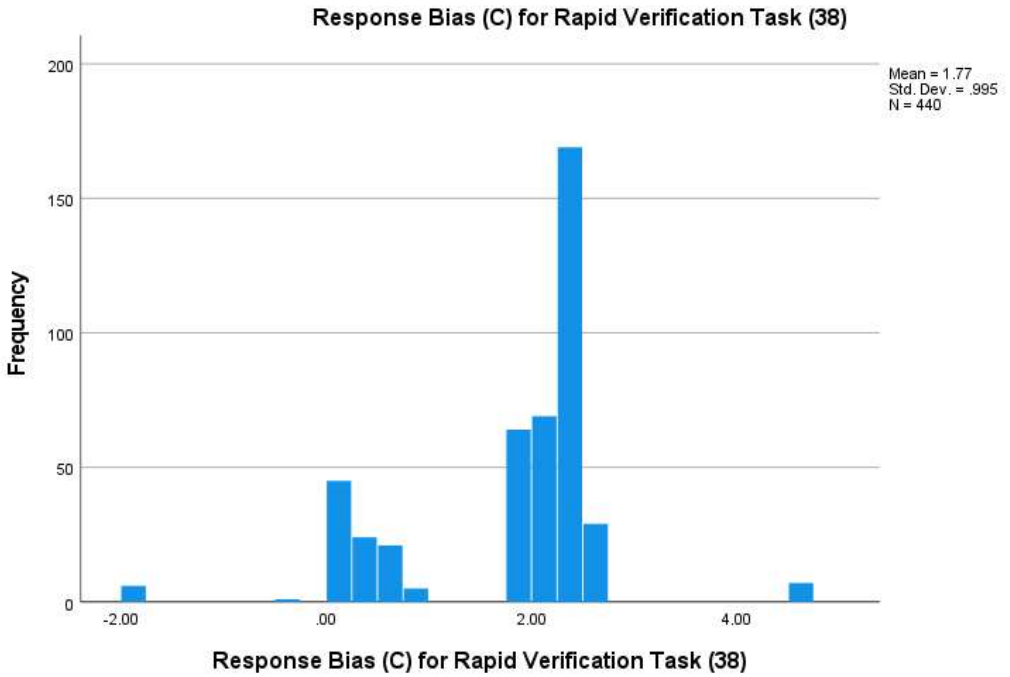


Figure 5.7

Histogram Showing Distribution of Response Bias for the Rapid Verification Task, Target 38



5.2.2 Whole Number Items – Constructing a Subscale

As in the Missing Symbol task, it was decided that binary items were unlikely to be useful indicators for an exploratory factor analysis. Sum scores were therefore calculated for each trial. The reliability for each trial was quite low, with a Cronbach's $\alpha = 0.37$ for target 59 and $\alpha = 0.47$ for target 38. This is likely to be at least partly a result of the fact that binary items produce minimal variation to be captured by inter-item correlations, as discussed previously. It is probably also due to the fact that the False item scores were extremely high while the True item scores were more varied. This would mean that the False items tended to behave differently to the True items, thus decreasing the internal consistency of the trial.

It was nonetheless decided to retain the False items, for several reasons. Firstly, excluding all False items would not leave many items for a sum score (particularly for target 59, where only four items out of 12 were True). Secondly, as per the discussion in Section 5.2.1, it seems likely that the differential response patterns on True and False items may reflect genuine and interestingly adaptive differences in strategy choice. Thirdly, as can be seen from Figures 5.8 and 5.9, there is at least a small amount of variation at the lower end of the score distribution that would be lost if all of the False items were omitted.

Figure 5.8

Histogram Showing Score Distribution for the Rapid Verification Task (Target 59)

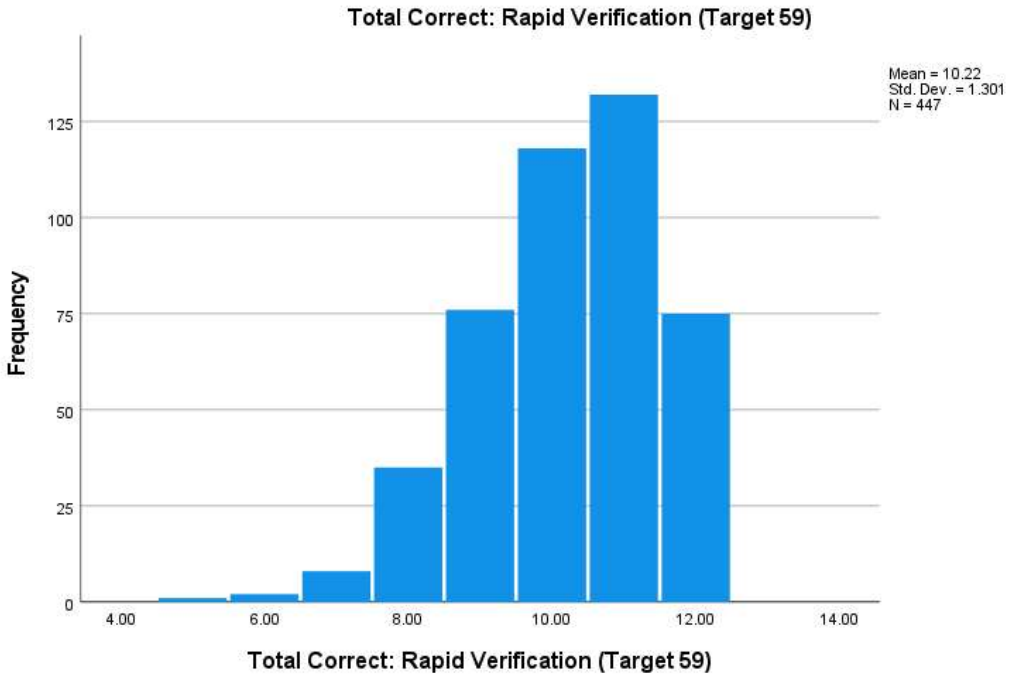
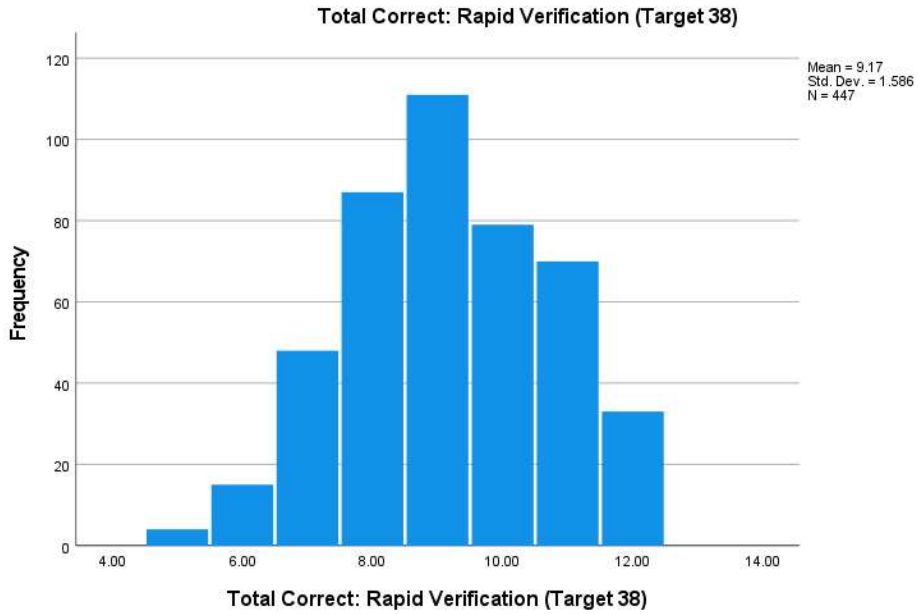


Figure 5.9

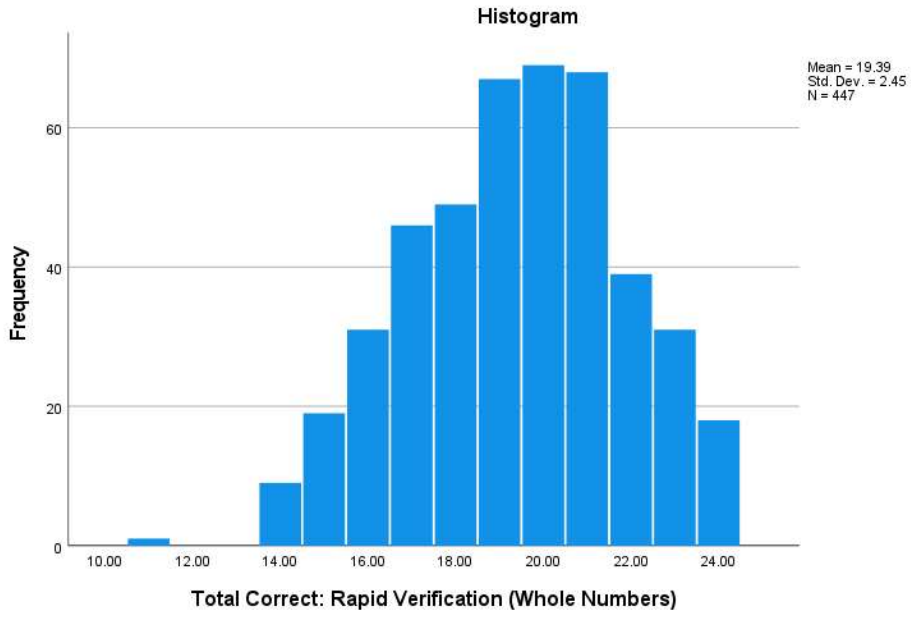
Histogram Showing Score Distribution for the Rapid Verification Task (Target 38)



However, it was decided to mitigate these issues by combining the sum scores for both trials into a two-item subscale for whole number trials. This subscale has a higher reliability at Cronbach's $\alpha = 0.60$ and a good amount of variability (as can be seen in Figure 5.10), albeit almost exclusively in the upper half of the possible range.

Figure 5.10

Histogram Showing Score Distribution for Rapid Verification Task (Whole Number Items)



5.2.3 Rational Number Items – Descriptive Statistics

Descriptive statistics for the rational number items in the Rapid Verification task are given in Table 5.7. The True items are shaded in green and the False items are shaded in yellow. The answers in the trial with target number ¾ followed a similar pattern to the whole number trials, with participants being extremely likely to answer False items correctly (means greater than 0.92), and less likely to answer True items correctly (means ranging from 0.33 to 0.82). A chi-square analysis confirmed that there was a very strong association between False items and correct answers, $\chi^2(1,447) = 461.07, p = 2.8 \times 10^{-10}$.

Table 5.7
Descriptive Statistics for the Rapid Verification Task (Rational Number Items)

	N	Mean	Std. Deviation	Skewness	Item Type
Rapid Verification Task ½ a	447	0.85	0.35	-2.02	False
Rapid Verification Task ½ b	447	0.89	0.32	-2.47	True
Rapid Verification Task ½ c	447	0.31	0.46	0.83	True
Rapid Verification Task ½ d	447	0.56	0.50	-.25	True
Rapid Verification Task ½ e	447	0.86	0.34	-2.13	True
Rapid Verification Task ½ f	447	0.90	0.30	-2.62	False
Rapid Verification Task ½ g	447	0.75	0.43	-1.18	True
Rapid Verification Task ½ h	447	0.49	0.50	0.05	False
Rapid Verification Task ¾ a	447	0.82	0.39	-1.64	True
Rapid Verification Task ¾ b	447	0.78	0.41	-1.38	True
Rapid Verification Task ¾ c	447	0.93	0.25	-3.40	False
Rapid Verification Task ¾ d	447	0.33	0.47	0.73	True
Rapid Verification Task ¾ e	447	0.92	0.27	-3.15	False
Rapid Verification Task ¾ f	447	1.00	0.07	-14.90	False
Rapid Verification Task ¾ g	447	0.41	0.49	0.36	True
Rapid Verification Task ¾ h	447	0.80	0.40	-1.47	True

However, in the trial with target ½, the difference between False and True items was not as clear. None of the False items had means exceeding 0.90, and while two of them had high means (0.85 and 0.90 for items ½a and ½f respectively), the third had a mean of only 0.49 (item ½h). The means for the True items ranged from 0.31 to 0.89, meaning that participants performed similarly well on the highest-scoring True items and False items. A chi-square

analysis still found a statistically significant association between False items and correct answers, $\chi^2(1,447) = 20.36, p = 6.4 \times 10^{-6}$, but it was not as marked as in the other three trials.

A probable explanation for this can be found in a closer examination of low-scoring items across both rational number trials. The lowest-scoring items by some margin are ½c ($\bar{x}$ = 0.31), ½h ($\bar{x}$ = 0.49), ¾d ($\bar{x}$ = 0.33) and ¾g ($\bar{x}$ = 0.41). All of these items involve division by a fraction, which was identified in Section 5.1.3 as something that participants seem to particularly struggle with. In the trial with target ¾, both items involving division were True, which meant that the tendency to answer them incorrectly worked in tandem with the tendency identified in Section 5.2.1 to answer True items incorrectly more frequently than False items. In the trial with target ½, however, one of the division items was False (item ½h). It seems that the difficulty of fraction division outweighed the relative ease of answering False items correctly; it also outweighed any possible non-circling bias.

However, the difficulty of fraction division cannot be regarded as a complete explanation of the overlapping answer patterns in the trial with target ½. For one thing, item ½a also involved division by a fraction, but it was answered correctly by 85% of participants. It is possible that in the case of this relatively simple expression $\left(0.5 \div \frac{1}{4}\right)$, the ease of ascertaining that the False item was in general not equal to ½ could have outweighed the added difficulty of fraction division. By contrast, in the case of ½h, $\left(\frac{3}{4} + \frac{1}{4}\right) \div \frac{1}{2}$, the expression's relative complexity – not to mention its similarity to $\left(\frac{3}{4} + \frac{1}{4}\right) \times \frac{1}{2}$, which would indeed have been equal to ½ – meant that the item's difficulty trumped the ease of rejecting False items.

Participants' response bias was investigated for each rational number trial to see if the same conservative strategy was apparent in these trials. Descriptive statistics for the response bias for each rational number trial are given in Table 5.8. The mean response bias is positive for both trials, indicating a conservative strategy, but in accordance with the less extreme response patterns in the trial with target ½, the response bias for that trial was considerably smaller than for the other three Rapid Verification trials. The mean response bias values indicated that the average threshold for circling items was 0.4 SD above the ideal observer's threshold for target ½, and 1.5 SD above the ideal observer's threshold for target ¾. The histogram in Figure 5.11 illustrates that while there was a mild positive response bias overall for target ½, over half of the participants had a *C*-score near zero, indicating a lack of response bias. The histogram in

Figure 5.12 illustrates that the much larger mean response bias for target ¾ was not merely a mathematical artefact: virtually no participants had a negative response bias in that trial.

Table 5.8

Descriptive Statistics for Response Bias (Rapid Verification Task, Rational Number Items)

	N	Min	Max	Mean	Std. Deviation	Skewness
Response bias (C) for Rapid Verification task ½	440	-4.50	4.50	0.41	1.42	0.20
Response bias (C) for Rapid Verification task ¾	446	-2.03	4.50	1.48	1.17	-0.09
Valid N (listwise)	**439**					

Figure 5.11

Histogram Showing Distribution of Response Bias for the Rapid Verification Task, Target ½

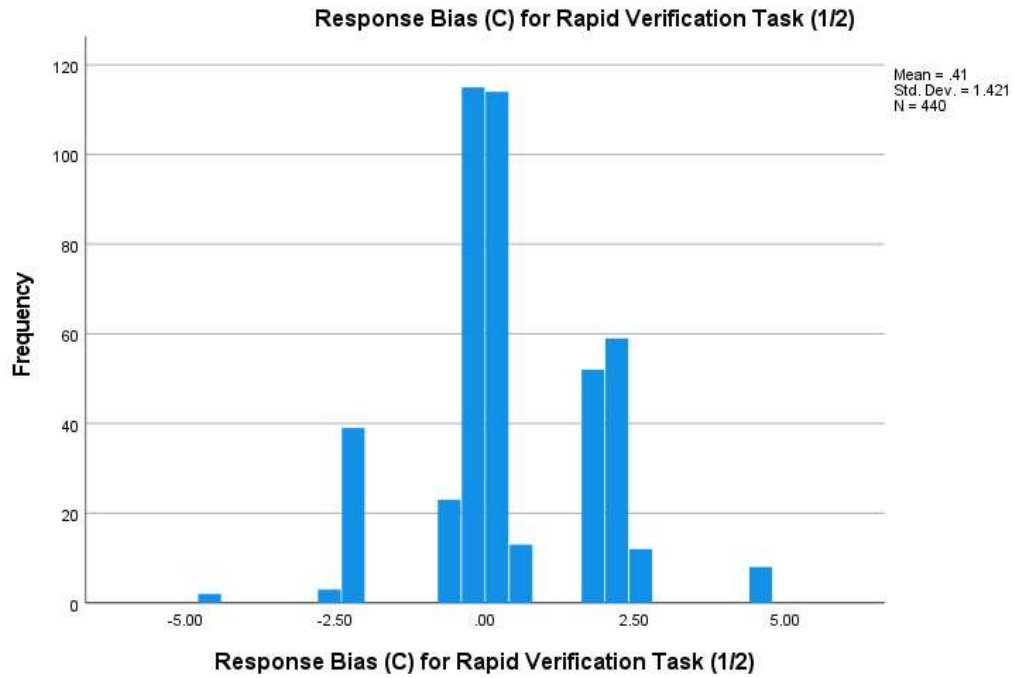
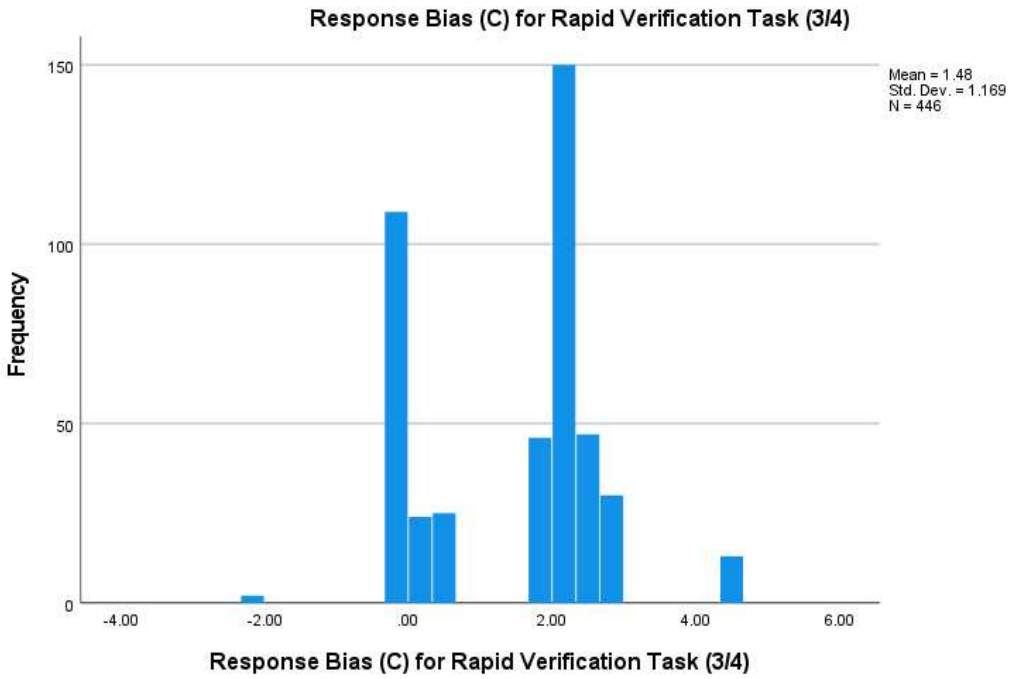


Figure 5.12

Histogram Showing Distribution of Response Bias for the Rapid Verification Task, Target ¾



5.2.4 Rational Number Items – Constructing a Subscale

In a similar fashion to the whole number Rapid Verification items, it was decided to combine the binary items into sum scores for each trial. As with the whole number items, the reliability for each trial was relatively low, which was unsurprising, given the binary nature of the items and the extremely high scores on some items. Cronbach's $\alpha = 0.38$ for target ½, and Cronbach's $\alpha = 0.60$ for target ¾.

Similarly to the whole number trials, it was decided to retain all items for further analysis. In the trial with target ½, no items had means above 0.9, meaning there was a fair amount of variation in each item. The score distribution for target ½ is given in Figure 5.13. In the trial with target ¾, several False items had means above 0.9, but the same logic was used as in Section 5.2.2: excluding False items would leave relatively few items, the differential response patterns between True and False items seemed to reflect meaningful differences in strategy choice, and excluding false items would mean losing some variation at the lower end of the score distribution (as can be seen in Figure 5.14). The sum scores for each trial were combined to make a two-item subscale for rational number trials. The rational number subscale had a

reliability of Cronbach's $\alpha = 0.71$ and a score distribution that can be seen in Figure 5.15. The score distribution exhibits a good amount of variation in scores in the upper two-thirds of the possible range.

Figure 5.13

Histogram Showing Score Distribution for the Rapid Verification Task (Target ½)

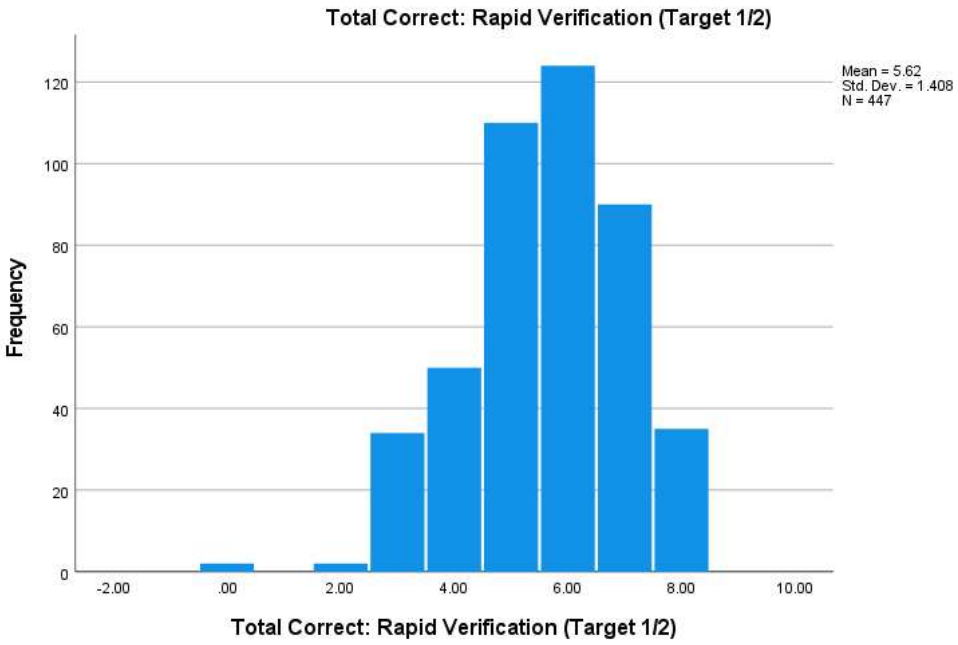


Figure 5.14

Histogram Showing Score Distribution for the Rapid Verification Task (Target ¾)

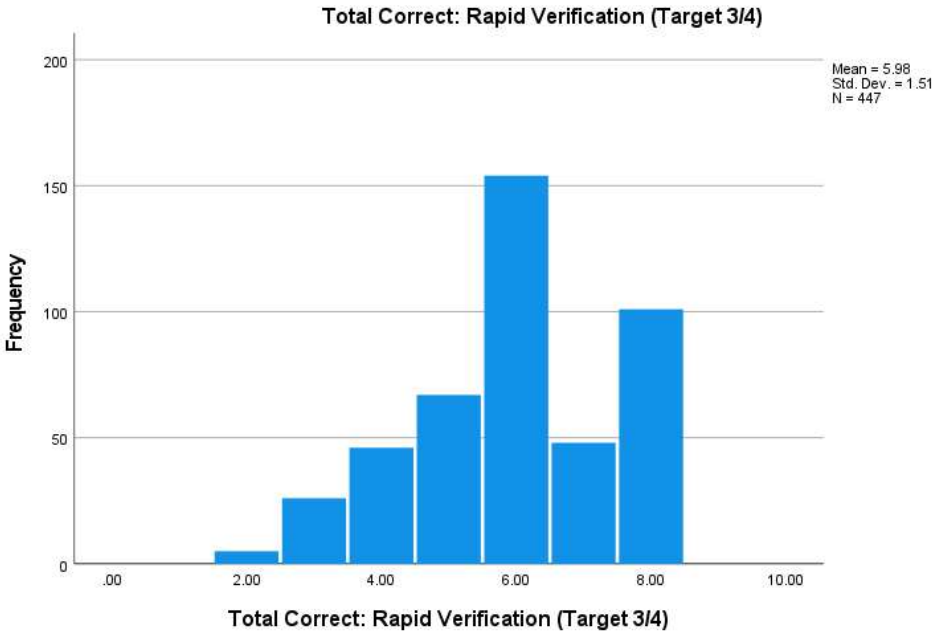
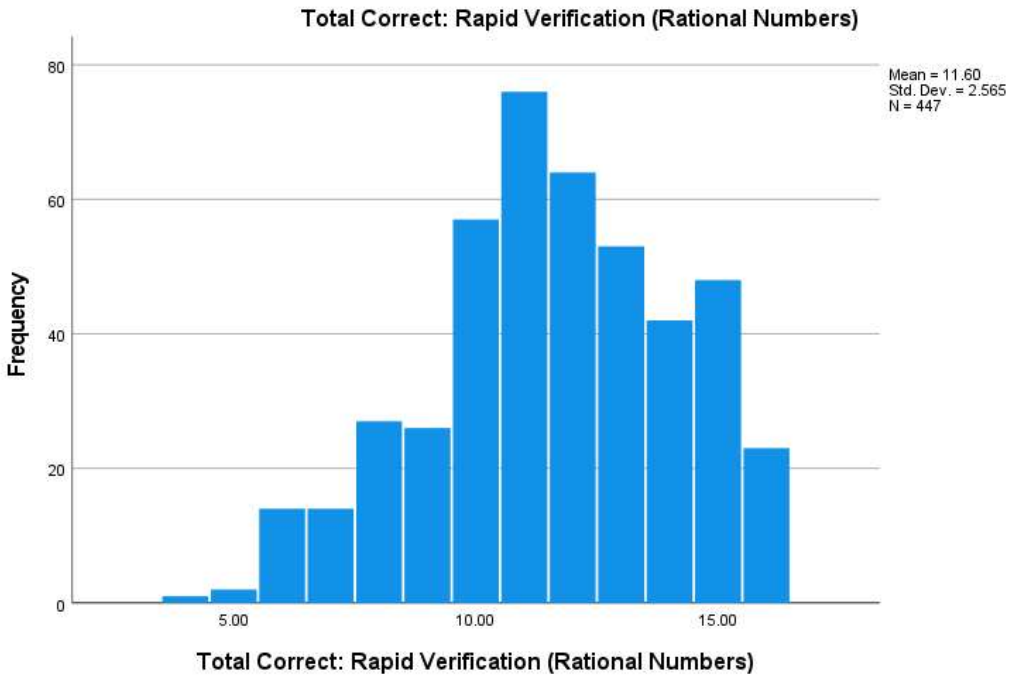


Figure 5.15

Histogram Showing Score Distribution for the Rapid Verification Task (Rational Number Items)



5.3 Exploratory Factor Analysis

This section reports the results of an exploratory factor analysis which investigated whether the three tasks measured a single unidimensional latent construct. Before proceeding with an EFA, it was necessary to conduct preliminary checks on the distributional properties of the data, sample size, and certain other aspects of the data, to establish whether the data were suitable for such an analysis. The results of the preliminary checks are also reported below.

Eight variables were considered in these analyses: (1 & 2) the number of correct responses on each of the two whole number items in the ASP task, (3 & 4) the number of correct responses on each of the two rational number items in the ASP task, (5) the number of correct responses on the revised whole number subscale in the Missing Symbol task, (6) the number of correct responses on the rational number subscale in the Missing Symbol task, (7) the number of correct responses on whole number items in the Rapid Verification task, and (8) the number of correct responses on rational number items in the Rapid Verification task.

The procedure used in this section is essentially identical to the procedure described in Chapter 3, so where an analysis or check has already been explained or justified in Chapter 3, the explanation or justification is omitted in this chapter. The analyses in this section are therefore described in a straightforward manner, with minimal referencing or explanation.

5.3.1 Preliminary checks on distributional properties of the data

Variability

There is no compelling reason to believe that the range of any of the variables of interest has been seriously attenuated by the sample selection process, even though a large majority of the participants were enrolled for Honors or AP courses. The ASP items, as noted in Section 3.2.1, demonstrate significant variability and no obvious ceiling or floor effects. Scores are also distributed throughout the range for both whole number and rational number subscales from the Missing Symbol task, without a clustering of values at the upper or lower ends of the scale (see Figures 5.4 and 5.5). In the Rapid Verification task subscales, the lowest values of the possible scoring range are not used at all (see Figures 5.10 and 5.15). This might suggest that some items in that task were too easy. However, the scores are approximately normally distributed across a fairly large range in the upper part of the possible scoring range for each subscale, which suggests that the task has nonetheless captured the full range of variability.

Linearity

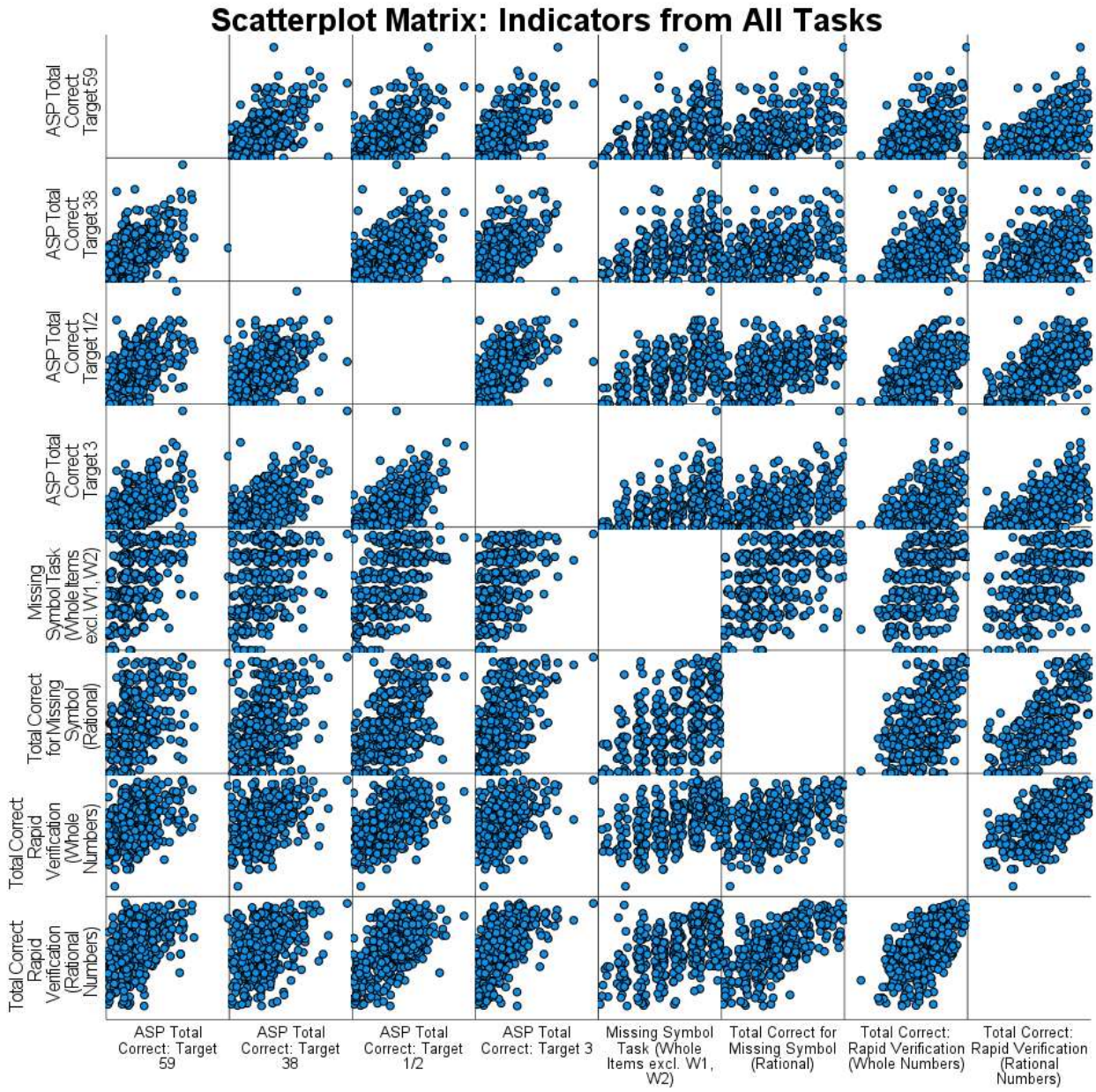
The linearity of the relation between each pair of indicator variables was first assessed by inspecting scatter plots. As in Chapter 3, the JITTER command was used to introduce a small amount of random noise, thus separating out coincident points and making the graphs easier to interpret. Jittered scatter plots comparing all eight variables are given in Figure 5.16.

Intra-task relationships: As observed in Section 3.2.1, there is a clear moderately positive linear relationship between each pair of variables from the ASP task. A similar moderately positive linear relationship can be seen between the two indicators from the Rapid Verification task. The scatter plot of the two Missing Symbol task indicators shows most clearly that

participants tended to score higher on the whole number items than the rational number items (as can be seen from the fact that low scores on the rational number items are associated with a wide range of scores on the whole number items, but not vice versa) – but it also shows a discernible, if fairly weak, positive linear relationship between the two variables.

Figure 5.16

Jittered Scatter Plots Showing the Relationship Between Indicators From All Tasks



Inter-task relationships: The ASP task indicators show a clear weak-to-moderate positive linear correlation with the Rapid Verification task indicators. The rational number Missing

Symbol task indicator also shows a moderate positive linear relationship with the Rapid Verification task indicators. The whole number Missing Symbol task indicator shows a weaker but still positive linear relationship with the Rapid Verification task indicators. The rational number Missing Symbol task indicator shows a clearly positive linear relationship with the ASP task indicators. The relationship appears to be stronger with the rational number ASP items (target ½ and target 3) than with the whole number ASP items (target 59 and target 38). The whole number Missing Symbol task indicator shows a less clearly linear relationship with the ASP task indicators. It seems that a high score on the whole number Missing Symbol task is a necessary but not sufficient condition to score highly on the ASP task. In other words, participants might score high on the whole number Missing Symbol task and low on the ASP task, but not vice versa. Based on this, one might expect weak positive correlations between these items.

Since the indicator variables generally appear to bear a linear relationship to each other, Pearson correlation coefficients were calculated, excluding cases listwise in the case of missing values. These coefficients are given in Table 5.9, and have been colour-coded as follows: 0.300-0.399: orange, 0.400-0.499: yellow, 0.500-0.599: light green, 0.600-0.699: dark green. The correlation coefficients ranged from $r = 0.374$ to $r = 0.686$, $p < 0.001$ for all Pearson coefficients. These moderate positive correlations between each pair of variables suggest that there is a fair amount of common variance which an EFA might be able to explain. None of the correlations are high enough to prompt worries about multicollinearity.

Table 5.9 also gives rise to an interesting observation that is unrelated to the requirements of an EFA. The correlations between rational number items (in the bottom-right corner of each block) tend to be stronger than the whole-whole correlations and the whole-rational correlations. This means that the rational number items have more in common with each other than other number type pairings do, even when they come from different tasks. There are at least two reasons that this might be the case. The first is that there is some aspect of adaptive number knowledge which is uniquely captured by the rational items and not by the whole items. One argument in favour of this could be that school students receive less instruction on and practice with rational numbers as compared to whole numbers – both in terms of time spent on the topic (since rational numbers are introduced several years after whole numbers, and typically only occupy a few weeks or months in the curriculum each year), and in terms of depth of mathematical explanation (since many teachers have poor rational number knowledge

themselves; Van Hoof et al., 2017). This might leave more room for individual differences in rational number sense and adaptive number knowledge to influence performance on rational number items, since students are less "coached" to perform well on them. In other words, it may be harder to design a whole number task that is genuinely novel for students, and which would therefore activate their adaptive number knowledge. If this conjecture is true, one might expect the rational number items to load more strongly on a latent factor of adaptive number knowledge (although this would by no means be conclusive evidence for the conjecture).

Table 5.9

Pearson Correlations Between All Eight Indicator Variables

	ASP Target 59	ASP Target 38	ASP Target ½	ASP Target 3	Missing Symbol revised whole subscale	Missing Symbol rational subscale	Rapid Verification (whole)	Rapid Verification (rational)
ASP Target 59	--							
ASP Target 38	0.55***	--						
ASP Target ½	0.53***	0.45***	--					
ASP Target 3	0.58***	0.54***	0.61***	--				
Missing Symbol revised whole subscale	0.42***	0.38***	0.52***	0.47***	--			
Missing Symbol rational subscale	0.44***	0.37***	0.51***	0.56***	0.46***	--		
Rapid Verification (whole)	0.39***	0.41***	0.48***	0.45***	0.50***	0.46***	--	
Rapid Verification (rational)	0.48***	0.40***	0.58***	0.57***	0.53***	0.69***	0.58***	--

*** Correlation is significant at the 0.001 level (2-tailed).

Listwise N=405

The second possible reason for the stronger correlations between rational number items is that there is some other aspect of rational arithmetic skills that is measured by all the rational number items, and which is causing the higher correlations. In this case, the increased

correlation would have nothing to do with the construct of interest, adaptive number knowledge. This is probably a more likely explanation, given that difficulty with fraction division has already been identified as a probable contributor to distinctive answer patterns in both the Missing Symbol task and the Rapid Verification task.

Normality

The normality of the ASP task items was found to be acceptable in Section 3.2.1. The normality of the Missing Symbol and Rapid Verification items was assessed in two ways. Firstly, histograms of the score distributions for each item were examined. The distribution for the revised whole number subscale of the Missing Symbol task (Figure 5.3) showed a definite skew to the left, but was not decisively non-normal. The distribution for the rational number subscale of the Missing Symbol task (Figure 5.5) looked close to normally distributed. The distribution for the whole number Rapid Verification items (Figure 5.10) also looked close to normally distributed, with the caveat that the lower end of the possible score range was not used at all. Finally, the distribution for the rational number Rapid Verification items (Figure 5.15) looked approximately normal, but with a more noticeable leftwards skew.

Based on the histograms, there appeared to be no reason for serious concern, but the skewness and kurtosis figures for each distribution were examined to get a more precise measurement of their normality. The figures, given in Table 5.10, confirm that skewness is well under 2.0 and kurtosis is well under 7.0.

Table 5.10

Descriptive Statistics for Missing Symbol Task & Rapid Verification Task Indicators

	N	Min	Max	Mean	Std. Deviation	Skewness	Kurtosis
Missing Symbol revised whole subscale	447	0	6	3.91	1.62	-0.46	-0.78
Missing Symbol rational subscale	447	0	7	3.13	1.87	0.34	-0.69
Rapid Verification (whole)	447	11	24	19.39	2.45	-0.21	-0.43
Rapid Verification (rational)	447	4	16	11.60	2.57	-0.30	-0.40

A final test of normality was conducted by means of the Kolmogorov-Smirnov test. The test indicated that the distributions of all eight indicator variables differed significantly from the normal distribution, as shown in Table 5.11. This result was not surprising, given the size of the present sample. However, since the histograms, the skewness figures and the kurtosis figures all indicated that the data were reasonably close to normally distributed, the results of the overly-sensitive Kolmogorov-Smirnov test do not change the conclusion that the data were suitable for EFA. (In fact, based on the foregoing checks, it was expected that the Kolmogorov-Smirnov test would not be necessary; however, it was conducted for the sake of consistency with the EFAs performed in Chapter 3.)

Table 5.11

Results of Kolmogorov-Smirnov Test

	Kolmogorov-Smirnov*		
	Statistic	df	Sig.
ASP Total Correct: Target 59	.214	405	.000
ASP Total Correct: Target 38	.156	405	.000
ASP Total Correct: Target 1/2	.097	405	.000
ASP Total Correct: Target 3	.157	405	.000
Missing Symbol revised whole subscale	.202	405	.000
Missing Symbol rational subscale	.158	405	.000
Rapid Verification (whole)	.106	405	.000
Rapid Verification (rational)	.088	405	.000
*Calculated using the Lilliefors Significance Correlation			

5.3.2 Preliminary check on sample size

As outlined in Section 3.2.2, the required sample size is affected by the ratio of variables to factors and the level of communality. The level of communality cannot be known in advance of conducting the EFA, but the variable-to-factor ratio can be calculated for the most likely outcomes of an EFA. The most likely outcomes of the present EFA would be:

- A one-factor solution: ANK underlies performance on all three tasks
- A two-factor solution:
    - ANK underlies performance on the ASP task, but routine knowledge underlies performance on the other two tasks, *or*

    - ANK underlies performance on the ASP task and one other task, but routine knowledge underlies performance on the third task, *or*
    - Performance on the rational number items is distinct from performance on the whole number items
- A three-factor solution: Performance on each task is explained by a different latent variable

Since there are eight indicator variables, a one-factor solution would have a variable-to-factor ratio of 8:1. In this case, Mundfrom et al. (2005) recommend a minimum sample size of 55 in the most demanding scenario (i.e. a low level of communality, excellent agreement between sample and population solutions). Therefore, the present sample size of $n = 405$ would be more than sufficient if it turned out there was one latent factor. A two-factor solution would have a variable-to-factor ratio of 4:1. In the most demanding scenario, Mundfrom et al. recommended a minimum sample size of 270 for a 4:1 ratio with a two-factor solution. Therefore, the present sample would again be sufficiently large. A three-factor solution would have a variable-to-factor ratio of 2.7:1. Mundfrom et al. only investigated scenarios with a minimum of three variables per factor. With a 3:1 variable-to-factor ratio and a four-factor solution, they recommend a minimum sample size of 1700 in the most demanding scenario. In this scenario, therefore, the present sample would not be large enough. However, if the level of communality were high and a merely good agreement between sample and population solutions were required (i.e. the least demanding scenario considered by Mundfrom et al.), a sample size of 260 would be sufficient. Even if the sample size were adjusted upward somewhat to allow for the fact that the ratio would be 2.7:1, the present sample size should be large enough in this scenario.

Therefore, the present sample is comfortably large enough if an EFA reveals a one-factor or two-factor solution. If a three-factor solution is revealed, the communalities should be examined carefully. If the communalities are lower than 0.6, the solution may not be trustworthy.

5.3.3 Preliminary checks on other aspects of the data

First, the data should be measured on a *ratio or interval scale*. All eight indicator variables are measured on an interval scale, so this is not problematic.

Second, the quantity and nature of any *missing values* should be considered. Overall, 42 participants (9.4%) were missing data for at least one of the indicator variables. As explained in Section 3.2.3, the missing data were generally a result of poor-quality scans and were unlikely to follow a systematic pattern. Therefore, cases with missing data were excluded listwise, leaving a sample size of 405.

Third, an inspection of *outliers* was conducted to ensure there were no out-of-range values or missing-value codes masquerading as real data points.

Fourth, the reliability of all indicator variables was examined. Cronbach's $\alpha = 0.882$ across the eight variables, which is well above the recommended minimum level of 0.70.

Finally, it is good to make a *general assessment* of whether the variables are correlated to the extent that suggests there are one or more latent factors. Subjectively, this can be done by examining the inter-item correlation matrix, as was done in Section 5.3.1. Objectively, Barlett's test of sphericity can test whether the correlation matrix is (un)likely to be generated by random data, and therefore whether an EFA is advisable. As explained in Section 3.2.3, Barlett's test of sphericity should be supplemented by the KMO measure of sampling adequacy when the sample size is large. Bartlett's test of sphericity indicated that the correlation matrix was significantly different to the identity matrix, $\chi^2(28, n = 405) = 1486.38, p > 0.001$. The KMO statistic was 0.91, which is classified as "marvellous" (Field, 2018, p. 798). The KMO values for individual indicators range from 0.868 to 0.938, all well above the acceptable level of 0.5. Taken together, these results provide further confirmation that the data are well-suited to factor analysis.

5.3.4 Results of the Exploratory Factor Analyses

As in Chapter 3, it was decided to conduct multiple analyses for the sake of comparison. This section reports the results of one PCA, and two EFAs using different factoring methods.

Principal Components Analysis

A PCA was conducted on the eight indicator variables. The communalities ranged from medium to fairly high, according to Mundfrom et al.'s (2005) definitions, ranging from 0.45-0.66. An initial analysis was run to obtain eigenvalues for each component in the data. Only one component had an eigenvalue that exceeded Kaiser's criterion of 1, and that component explained 56.13% of the observed variance. The scree plot clearly confirmed the existence of a single major component. As only one component was identified, the solution was not rotated. Table 5.12 shows that all eight variables loaded strongly onto the identified component, with loadings ranging from 0.67 to 0.81.

Table 5.12

Component/Factor Matrices for Three Different Extraction Methods

Indicator	**Loading on Component 1 (PCA)***	**Loading on Factor 1 (PAF)***	**Loading on Factor 1 (ML)***
Rapid Verification (rational)	0.81	0.79	0.79
ASP Target 3	0.81	0.78	0.77
ASP Target ½	0.79	0.75	0.75
Missing Symbol rational subscale	0.76	0.71	0.73
ASP Target 59	0.73	0.68	0.68
Missing Symbol revised whole subscale	0.71	0.66	0.66
Rapid Verification (whole)	0.71	0.65	0.65
ASP Target 38	0.64	0.61	0.61

*PCA = Principal Components Analysis, PAF = Principal Axis Factoring, ML = Maximum Likelihood

Exploratory Factor Analyses

Two EFAs were conducted on the eight indicator variables, one using principal axis factoring and the second using maximum likelihood estimation. The communalities after extraction ranged from 0.38-0.62 in the case of principal axis factoring, and 0.37-0.63 for maximum

likelihood estimation. The initial analyses were identical to the initial analysis for the PCA reported above, with a single factor with an eigenvalue greater than 1 explaining 56.13% of the variation. After extraction, this major factor explained 50.03% (principal axis factoring) or 50.02% (maximum likelihood) of the variation. The scree plots clearly confirmed an interpretation with a single major component. As only one factor was identified, the solution was not rotated. Table 5.12 shows the factor loadings for both EFAs. The loadings are slightly lower than in the PCA, but still high, ranging from 0.61-0.79 with principal axis factoring and 0.61-0.79 with maximum likelihood estimation.

5.4 Discussion and Conclusions

The results of the exploratory factor analyses provide good evidence that the test is unidimensional. The latent factor explains over 50% of the variance in test scores, and all indicators are strongly associated with the factor, with loadings of over 0.6 for every indicator. The choice of a PCA or EFA does not affect this result, which was consistent across all three methods of extraction. This is an important first step towards developing a multi-faceted measure of ANK which could give a more precise understanding of the construct.

The question, of course, is *whether the latent factor is actually adaptive number knowledge*. It is possible that it is. The ASP task is an established measure of adaptive number knowledge, and this chapter has found that the other tasks in this test measure a factor which is also measured by the ASP items. Therefore, it is possible that all tasks are measuring the same latent variable of adaptive number knowledge. With the present data, the question cannot be decisively determined, but it can at least be concluded with some degree of confidence that the three test items do not measure different things (i.e. these tasks have convergent validity). Further research would need to investigate discriminant validity. Since the tasks are timed, it is possible that the latent factor is actually a measure of processing speed or the absence of test anxiety. The latent factor could also be general mathematical ability. Finally, a middle-ground interpretation may also be true – the test may reflect both adaptive number knowledge and general mathematical ability. It could be that one or the other acts as a mediating variable: ANK might support the development of strong general mathematical ability, which in turn results in high scores on the present test instrument; or strong general mathematical ability might contribute to high levels of ANK, which are detected by the present test instrument. Or

it could be that high levels of ANK are inextricably bound up with general mathematical ability, that ANK is a part of mathematical ability, rather than a cause or effect of it.

If one assumes that the present three-task test does measure ANK, the next question would be *whether it contributes to a better understanding of the construct*. This is a pertinent question since the three-task test seems to generate some extra noise when compared to the ASP items alone. This can be seen from the fact that the factor loadings are lower in the three-task EFAs conducted in this chapter than in the EFA conducted exclusively with ASP items in Chapter 3. Using the PCA loadings as an illustration, the average factor loading in the ASP-only analysis was 0.81, while in the three-task analysis it was 0.75 – still a high average, but considerably lower than in the ASP-only test.

Despite the added noise, the three-part test still provides valuable insights into the nature of adaptive number knowledge. For one thing, the results imply that ANK plays an important role not only in creative processes, such as the open-ended ASP task, but also in responsive processes where the task is more constrained, such as the Missing Symbol and Rapid Verification tasks. This in turn would suggest that ANK is an important building block not only of "blue-sky" mathematical invention, but also the type of inventive thinking that is more often required in daily mathematical applications, where one must work within given constraints. Secondly, a small drop in factor loadings is to be expected when moving from a single-task EFA to a multi-task EFA. This is because a single-task EFA will detect common variance which is due to both the underlying construct of interest and generic task characteristics, but as task characteristics differ across items in a multi-task EFA, the latter analysis is less likely to pick up variance that is due to task characteristics but not to the construct of interest. For instance, a student who had high ANK but did not realise that commutative variations like $2 \times \frac{1}{4}$ and $\frac{1}{4} \times 2$ would count as two different and valid solutions may have generated fewer solutions in the ASP task than a student with similarly high ANK who utilised such variations. This variation is due to interpretation of the task instructions, and not to the level of ANK. The first student would likely perform as well as the second student on the Missing Symbol and Rapid Verification tasks, where the expectations were clearer. This would lead to slightly different patterns of responses across the ASP task and the other tasks, potentially decreasing the factor loadings, but does not necessarily have any bearing on how well the three-task test measures ANK. In fact, the three-task test might measure ANK more accurately, because it

would be less likely to over- or under-estimate a person's level of ANK as a result of task characteristics. In this view, the higher loadings in the ASP-only model might actually reflect *more* noise in that model than the three-task model, since the higher loadings might be caused by task characteristics (noise) rather than ANK (signal).

Knowing that ANK plays a role in different mathematical situations is valuable. However, there is probably some noise in the test items which future research could aim to reduce. Potential problems that were raised earlier in this chapter, such as the uneven distribution of fraction division questions and the ceiling effect on whole number False items in the Rapid Verification task, would be sensible places to focus in order to improve the validity of the present test.

A third question which could be asked is *why do the rational number indicators load more heavily onto the latent factor than the whole number indicators?* In all three analyses (PCA, principal axis factoring and maximum likelihood extraction), the four indicators with the strongest loadings were the rational indicators. One possible answer is that there was more variation in the rational number scores: in both the Missing Symbol task and (especially) the Rapid Verification task, scores for the whole number items tended to be higher than for the rational number items. This left less room for variation, and therefore less room for common variance to be detected in these items. Therefore, the rational number indicators may have been more decisive in the analysis simply because they varied more. A second possible answer may be that the rational number items were more novel to the students and therefore better at eliciting a response based on ANK rather than on routine expertise, as suggested in Section 5.3.1.

In conclusion, the exploratory factor analyses conducted in this chapter indicated that the three-task test was unidimensional, providing support for the hypothesis that all three items may measure adaptive number knowledge. If this is true, it suggests that ANK plays an important role in responsive mathematical thinking as well as creative mathematical thinking. As the current results only imply convergent validity, more research would be necessary to ascertain divergent validity. In addition, further iterations of this test could improve test validity by focusing on issues such as the uneven distribution of fraction division questions, the need for more comparable items in the whole number and rational number items, the ceiling effect on

False items in the Rapid Verification task, and requiring participants to indicate clearly whether they have considered each item in the Rapid Verification task.

6. Conclusion

This thesis set out to answer three research questions, namely:

1. Are adaptive whole number knowledge and adaptive rational number knowledge empirically distinct constructs?
2. Are there quantitative and/or qualitative individual differences in high school students' ANK and ARNK?
3. Does an instrument consisting of three different tasks provide a more nuanced understanding of A(R)NK than the ASP task provides alone?

This chapter will summarise the findings that answer these three research questions, highlighting new contributions to the literature on A(R)NK. It will then discuss the limitations of the present study, and suggest directions for future research.

6.1 Contributions to the Literature

In order to answer the *first research question*, an exploratory factor analysis was conducted to test whether adaptive whole number knowledge and adaptive rational number knowledge could be distinguished by the sample's scoring patterns on the arithmetic sentence production task. The answer appeared to be no: a single latent factor was identified, suggesting that the two types of ANK could be understood as a single construct. This is a new contribution to the literature on adaptive number knowledge, as adaptive whole number knowledge has never before been directly compared to adaptive rational number knowledge. It is an interesting finding: since rational number arithmetic requires many skills and concepts that whole number arithmetic does not, and which many students never fully master, it would have been plausible that some people would have well-developed whole number ANK but poor levels of ARNK. The fact that this does not seem to be the case strengthens the conception of high-level ANK as something that is characteristic of mathematical high-achievers. High-achieving students are more likely to develop the richly connected knowledge of relations between whole numbers that is ANK, and it seems probable that students who have this strong foundational framework would be well-positioned to extend it to include rational numbers as well, once they become familiar with them.

In order to answer the *second research question*, the results of the ASP task were analysed through ANOVAs, independent samples t-tests and a TwoStep cluster analysis. The cluster analysis produced a five-cluster model which bears substantial resemblance to the models found in earlier studies (McMullen et al., 2017, 2019). However, the five-cluster model in this study also revealed a split between middle-achieving and low-achieving students, which was not found in earlier studies. It is significant that a similar structure was found in this sample of American high schoolers, because previous large-scale latent group analyses have been conducted exclusively on younger students in Finland. The distinction between middle- and low-achievers might reflect a widening achievement gap or disinterest in the test that is specific to older students. Should this finding be replicated by future studies, it could provide interesting insights into how ANK develops (or fails to develop) in older students.

A sub-question of the second research question was: *Are there systematic differences between students in terms of age or school grade?* No difference was found between students of different ages in the total number of correct solutions generated, the number of complex solutions generated, or the number of cross-notational solutions generated, and minimal difference was found between students from different school grades. This is significant, because a clear correlation between age, school grade and scores on the ASP task is apparent in the studies with Finnish elementary and middle-schoolers (McMullen et al., 2017, 2019). The absence of age- or grade-based difference suggests that ANK growth may level off in the teen years. Alternatively, the discrepancy may be due to differences between Finland and the USA. One likely source of difference is that students in Finnish elementary and lower secondary schools progress through mathematics content at the same rate as their peers of the same age and grade, while in US high schools students in any given grade may enrol in a range of different mathematics modules. Although age was not found to be related to outcomes on the ASP task, a strong relationship was found between ASP outcomes and mathematics module enrolment, suggesting that mathematical experience may have a greater bearing on ANK levels than age does. This finding also confirms that high ANK is a correlate of high mathematical achievement. A strong relationship was also found between ASP outcomes and gender, with boys being significantly more likely than girls to score highly, generate complex solutions, and generate cross-notational solutions. Gender was independent of mathematics module enrolment, so this is a distinct finding, and one that is completely novel in A(R)NK research to date.

In order to answer the third research question, an exploratory factor analysis was conducted to assess whether the three tasks could all be considered to measure adaptive number knowledge. The EFA identified a single latent factor, suggesting that the test was unidimensional. While the analysis conducted in this study cannot confirm that the latent factor is definitely adaptive number knowledge, it is a first step towards establishing whether this test successfully measures ANK in a more multi-faceted way. If the latent factor is indeed adaptive number knowledge, then this would indicate that ANK plays an important role in responsive processes – as typified by the Missing Symbol and Rapid Verification tasks – as well as creative processes like the established ASP task. In the course of analysing the three-task test, this thesis also identified certain types of questions that seemed to be particularly easy (e.g. two-step whole number equations in the Missing Symbol task, False statements in the Rapid Verification task), particularly difficult (e.g. those involving fraction division, those Missing Symbol items requiring both a number and an operator to be filled in), or lacking clarity (e.g. whether uncircled False items in the Rapid Verification task had been rejected or skipped). This analysis may be useful in designing future iterations of this test.

6.2 Limitations of the Present Study

As with many studies, the sample limits the generalisability of the findings. Data was collected in only one school, so it may not be an accurate representation of ANK in US (or even south-eastern US) schools. In addition, all ANK research to date has been conducted exclusively in Finland and the USA. This may also limit its generalisability. Prior research has found that textbook approach influences adaptive strategy choice in mathematics (Sievert et al., 2019) and that national cultural differences have a strong impact on whether students develop a reflective and adaptive disposition towards academic work (Hess & Azuma, 1991), so it seems likely that both local and national context could affect the development of adaptive number knowledge. Finally, the present sample included a high percentage of students who were enrolled in Honors and AP classes, which may have skewed the sample towards higher achievers. This means that the findings of this study should be generalised to lower-achieving students (potentially more "normal-achieving" students) with caution.

The conclusions that could be drawn were limited by certain elements of task design. In the ASP task, the whole number items were sparse, while the rational number items were relatively dense. Dense and sparse items inherently attract different answer patterns, in terms of number of responses as well as whether responses are complex or simple. This meant that the whole number and rational number items were not easily comparable; any differences detected, like the lower proportion of complex solutions on rational number items, could not confidently be ascribed to either item type (dense/sparse) or number type (whole/rational). It could also potentially have obscured a genuine difference in adaptive number knowledge on whole and rational items, although this is perhaps not too likely, given that the EFA encompassing all three tasks also found a single factor across whole and rational items.

Continuing with limitations stemming from task design, the inclusion of fraction division in the Missing Symbol and Rapid Verification tasks seemed to be a powerful predictor, potentially even swamping the effect of adaptive number knowledge. Since fraction division tasks were not evenly distributed across True and False question types in the Rapid Verification task, this restricted the extent to which differences could be ascribed to question type, adaptive number knowledge, or simply the presence of fraction division. A third limitation stemming from task design is that, because it was unclear whether uncircled items in the Rapid Verification task had been rejected or skipped, interpretations based on the assumption that they were actively rejected should be read cautiously.

The low variability stemming from a ceiling effect in most of the False items in the Rapid Verification task may have influenced inter-item correlations and therefore the outcomes of the exploratory factor analysis in Chapter 5. Finally, the exploratory factor analyses conducted in Chapters 3 and 5 lack discriminant validity, because no potentially confounding variables were tested. It therefore cannot be established whether the latent factors identified by the EFA represent adaptive number knowledge, or some other construct like arithmetic fluency, routine conceptual knowledge, processing speed or test anxiety.

6.3 Recommendations for Future Research

Adaptive (rational) number knowledge is an area of scholarship where more studies are needed in general. Very few A(R)NK studies have been carried out to date, so it is impossible to draw

conclusions with any degree of certainty. In line with Section 6.2, there would be room to expand on the present study, specifically in terms of (a) investigating whether the latent factors identified in the two EFAs actually represent ANK or some other construct, and (b) improving the task design in the Missing Symbol and Rapid Verification tasks to make interpretation and comparison easier. Relatively minor changes in task design could make a significant difference, like including dense whole number items and sparse rational number items in the ASP task, distributing fraction division tasks evenly across True and False items in the Rapid Verification task, and requiring participants to actively write something to indicate rejection of an expression in the Rapid Verification task. Other aspects of task design could also be interrogated, such as the difficulty and usefulness of different types of Missing Symbol items (e.g. two-step equations, three-step equations, equations with missing numbers/operations/both), whether True items in the Rapid Verification task have higher discriminatory power than False items, and whether rational number items have a unique ability to elicit information about ANK (as suggested in Section 5.3.1).

It would also be instructive to further probe the developmental trajectory of A(R)NK. One way to do this would be to investigate the directionality of the relationship between mathematics module enrolment and ANK: does participation in higher-level mathematics courses improve ANK by requiring students to apply and integrate their existing arithmetic knowledge in new ways, or do these courses simply attract high achievers who happen to have higher ANK already? Or is the relationship bidirectional and mutually reinforcing? Another way to investigate the developmental trajectory of A(R)NK would be to examine whether adaptive whole number knowledge and adaptive rational number knowledge diverge in samples of different abilities or different ages to the present sample. In particular, it would be interesting to test whether the two constructs converge for students who are just beginning to learn about rational numbers. It seems plausible that such students may have already well-developed whole number ANK, but that a correspondingly high level of ARNK would only develop with extended exposure to rational numbers. A third way to examine the developmental trajectory of A(R)NK would be to conduct a person-centred analysis (such as a cluster analysis or latent profile analysis) using the same indicators from all three tasks that were used in the EFA in Chapter 5. This might reveal that high performance in some tasks is a prerequisite for high performance on other tasks, much like McMullen et al. (2020) found that routine procedural knowledge of rational numbers was a prerequisite for routine conceptual knowledge, and that both were prerequisites for ARNK as measured by the ASP task.

Another important topic for further investigation is the relationship between gender and ANK. Future research could investigate the robustness of this finding in different samples, and if it is robust could further investigate the reasons as to why it might be the case. It is important to establish whether the ASP task is a valid measure of inter-gender differences, or whether the apparent gender effects are caused by task characteristics (e.g. a timed test might be more likely to trigger an anxiety response in girls than in boys, perhaps the competitive tone set by the instruction "make as many as you can" appeals more to boys than girls, etc.). If the instrument is valid, it would then be important to examine why these gender differences exist, and whether the reasons are neurological, sociocultural, or both. It would be particularly interesting to examine whether the interaction between gender and adaptive mathematical expertise is consistent across countries, since PISA mathematics test results show that boys outperform girls in some countries (like the USA) but not others (like Finland) (OECD, 2019a, 2019b).

Finally, a qualitative analysis of all solutions generated in the ASP task could be fruitful. Examining notationally different but mathematically equivalent solutions may provide more insight into whether students understand the connection between fractions and decimals. For instance, a student who does not combine fractions and decimals in the same answer, but repeatedly swaps out fractions for decimals in consecutive solutions (e.g. 1.5×2 followed by $\frac{3}{2} \times 2$) would seem to understand the equivalence of these representations, although they would not have been given credit by a purely quantitative count of cross-notational solutions. A qualitative analysis of individual solutions would also enable an investigation of whether participants who did poorly in fraction division items in Missing Symbol and Rapid Verification tasks were able to avoid fraction division in the ASP task. If the answer is yes, this would help to explain why participants performed relatively better on the rational number items in the ASP task, but not in the Missing Symbol or Rapid Verification tasks.

6.4 Conclusion

This thesis has examined the performance of a sample of 447 Grade 9-12 students on three different measures of adaptive number knowledge with both whole numbers and rational numbers. The point of departure was novel in several ways: this is the first research project to investigate ANK in students of this age, to compare adaptive whole and rational number

knowledge directly, and to evaluate new ways of measuring ANK beyond the arithmetic sentence production task. The research produced three major findings. Firstly, adaptive whole number knowledge and adaptive rational number knowledge are not distinct in this sample; they seem to behave as a single construct. Secondly, participant performance on the ASP task can be described by a five-cluster model, and membership of high-achieving clusters is strongly and distinctly associated with both mathematics module enrolment and gender, but not with age or school grade. Thirdly, the test comprising of the ASP task and two new tasks is unidimensional, which suggests that all three tasks may measure adaptive number knowledge, although all three tasks could be improved to make them more easily comparable and interpretable. The three-task test is therefore a strong candidate for further evaluation and refinement in the quest for a more comprehensive measure of adaptive number knowledge.

References

Abdi, H. (2010). Signal Detection Theory. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 407-410). Elsevier Science.

ACER [Australian Council for Education Research]. (n.d.) *PISA and TIMSS* [Fact sheet]. https://www.acer.org/files/TIMSSandPISA-backgrounder.pdf

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis*. Sage.

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*(5), 775-785. https://doi.org/10.1111/desc.12155

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The Development of Arithmetic Concepts and Skills: Constructing Adaptive Expertise* (pp. 1-33). Lawrence Erlbaum Associates.

Benassi, M., Garofalo, S., Ambrosini, S., Sant'Angelo, R. P., Raggini, R., De Paoli, G., Ravani, C., Giovagnoli, S., Orsoni, M., & Piraccini, G. (2020). Using two-step cluster analysis and latent class cluster analysis to classify the cognitive heterogeneity of cross-diagnostic psychiatric inpatients. *Frontiers in Psychology, 11*, Article 1085. https://doi.org/10.3389/fpsyg.2020.01085

Boaler, J. (1998). Open and closed mathematics: Student experiences and understandings. *Journal for Research in Mathematics Education, 29*(1), 41-62. https://doi.org/10.5951/jresematheduc.29.1.0041

Brezovsky, B., McMullen, J., Veermans, K., Hannula-Sormunen, M. M., Rodríguez-Aflecht, G., Pongsakdi, N., Laakkonen, E., & Lehtinen, E. (2019). Effects of a mathematics game-based learning environment on primary school students' adaptive number knowledge. *Computers & Education, 128*, 63-74. https://doi.org/10.1016/j.compedu.2018.09.011

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3*(1), 21-29. https://doi.org/10.1111/j.2044-835X.1985.tb00951.x

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors

across two ECLS-K cohorts. *AERA Open, 2*(4), 1-19. https://doi.org/10.1177/2332858416673617

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology, 10*(3), 329-358. https://doi.org/10.1348/135910705X25697

College Board. (2019). *AP Program Summary Report 2019.* https://reports.collegeboard.org/media/pdf/Program-Summary-Report-2019_1.pdf

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29. https://doi.org/10.1037/1082-989X.1.1.16

Data Commons. (2022). *United States of America: Population enrolled in Grade 12.* https://datacommons.org/tools/timeline#place=country%2FUSA&statsVar=Count_Person_DetailedEnrolledInGrade12

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology, 7*, 1-16. https://doi.org/10.3389/fpsyg.2016.00508

Elia, I., Van den Heuvel-Panhuizen, M., & Kolovou, A. (2009). Exploring strategy use and strategy flexibility in non-routine problem solving by primary school high achievers in mathematics. *ZDM – Mathematics Education, 41*(5), 605-618. https://doi.org/10.1007/s11858-009-0184-6

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). SAGE.

Field, A. (2018). *Discovering Statistics Using SPSS* (5th ed.). SAGE.

Finnish National Agency for Education. (2018). *Compulsory Education in Finland.* https://www.oph.fi/sites/default/files/documents/compulsory_education_in_finland.pdf

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*(3), 165-190. https://doi.org/10.1006/jecp.1999.2532

Ganley, C. (2018). Are boys better than girls at math? *Scientific American.* https://www.scientificamerican.com/article/are-boys-better-than-girls-at-math/

Gaskin, J. (2012, March 20). *Two-step cluster analysis in SPSS* [Video]. YouTube. https://youtu.be/DpucueFsigA

Gaskin, J. (2015, July 2). *Validating a two-step cluster analysis – how many clusters?* [Video]. YouTube. https://youtu.be/Odk0kLuUGvY

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment, 27*(3), 265-279. https://doi.org/10.1177/0734282908330592

Gelbard, R., Goldman, R., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering, 63*(1), 155-166. https://doi.org/10.1016/j.datak.2007.01.002

Graven, M., Venkat, H., Westaway, L., & Tshesane, H. (2013). Place value without number sense: Exploring the need for mental mathematical skills assessment within the Annual National Assessments. South African Journal of Childhood Education, 3(2), 131-143. https://doi.org/10.4102/sajce.v3i2.45

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

Hatano, G. (1982). Cognitive consequences of practice in culture specific procedural skills. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 4*(1), 15-18. http://www.lchc.ucsd.edu/Histarch/newsletters.html

Hatano, G. (2003). Foreword. In A. J. Baroody & A. Dowker (Eds.), *The Development of Arithmetic Concepts and Skills: Constructing Adaptive Expertise* (pp. xi-xiii). Lawrence Erlbaum Associates.

Hatano, G., & Inagaki, K. (1984). Two courses of expertise. *Research and Clinical Center for Child Development Annual Report, 6*, 27-36. http://hdl.handle.net/2115/25206

Hecht, S. A., & Vagi, K. J. (2010). Sources of group and individual differences in emerging fraction skills. *Journal of Educational Psychology, 102*(4), 843-859. https://doi.org/10.1037/a0019824

Heinze, A., Star, J. R., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM – Mathematics Education, 41*(5), 535-540. https://doi.org/10.1007/s11858-009-0214-4

Hess, R. D., & Azuma, H. (1991). Cultural support for schooling: Contrasts between Japan and the United States. *Educational Researcher, 20*(9), 2-8 & 12. https://doi.org/10.3102%2F0013189X020009002

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class,

latent profile, and latent transition analysis. *Learning and Individual Differences, 66,* 4-15. http://dx.doi.org/10.1016/j.lindif.2017.11.001

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139-155. https://doi.org/10.1037/0033-2909.107.2.139

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterise math performance. *Science, 321*(5888), 494-495. https://doi.org/10.1126/science.1160364

IBM Corporation. (2021a). *Choosing a Procedure for Clustering*. https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=features-choosing-procedure-clustering

IBM Corporation. (2021b). *KMO and Bartlett's Test*. https://www.ibm.com/docs/en/spss/27.0.0?topic=detection-kmo-bartletts-test

IBM Corporation. (2021c). *Model Summary View*. https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=viewer-model-summary-view

IBM Corporation. (2021d). *SCATTER Subcommand (IGRAPH command)*. https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=command-scatter-subcommand-igraph

Kärki, T., McMullen, J., & Lehtinen, E. (2022). Improving rational number knowledge using the NanoRoboMath digital game. *Educational Studies in Mathematics, 110*(1), 101-123. https://doi.org/10.1007/s10649-021-10120-6

Kent, P., Jensen, R. K., & Kongsted, A. (2014). A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data: SPSS TwoStep Cluster analysis, Latent Gold and SNOB. *BMC Medical Research Methodology, 14*, Article 113. https://doi.org/10.1186/1471-2288-14-113

Lortie-Forgues, H., Tian, J., & Siegler, R. S. (2015). Why is learning fraction and decimal arithmetic so difficult? *Developmental Review, 38*, 201-221. http://dx.doi.org/10.1016/j.dr.2015.07.008

MacMillan, N. A. (2002). Signal Detection Theory. In H. Pashler & J. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology, Volume 4: Methodology in Experimental Psychology* (3rd ed., pp. 43-90). John Wiley & Sons.

Mazzocco, M. M. M., & Devlin, K. T. (2008). Parts and 'holes': gaps in rational number sense among children with vs. without mathematical learning disabilities. *Developmental Science, 11*(5), 681-691. https://doi.org/10.1111/j.1467-7687.2008.00717.x

McMullen, J., Brezovszky, B., Hannula-Sormunen, M., Veermans, K., Rodríguez-Aflecht, G., Pongsakdi, N., & Lehtinen, E. (2017). Adaptive number knowledge and its relation to arithmetic and pre-algebra knowledge. *Learning and Instruction, 49*, 178-187. https://doi.org/10.1016/j.learninstruc.2017.02.001

McMullen, J., Brezovszky, B., Rodríguez-Aflecht, G., Pongsakdi, N., Hannula-Sormunen, M., & Lehtinen, E. (2016). Adaptive number knowledge: Exploring the foundations of adaptivity with whole-number arithmetic. *Learning and Individual Differences, 47*, 172-181. https://doi.org/10.1016/j.lindif.2016.02.007

McMullen, J., Hannula-Sormunen, M. M., Lehtinen, E., & Siegler, R. S. (2020). Distinguishing adaptive from routine expertise with rational number arithmetic. *Learning and Instruction 68*, Article 101347. https://doi.org/10.1016/j.learninstruc.2020.101347

McMullen, J., Hannula-Sormunen, M. M., Lehtinen, E., & Siegler, R. S. (2022). Predicting adaptive expertise with rational number knowledge. *British Journal of Educational Psychology, 92*(2), 688-706. https://doi.org/10.1111/bjep.12471

McMullen, J., Kanerva, K., Lehtinen, E., Hannula-Sormunen, M. M., & Kiuru, N. (2019). Adaptive number knowledge in secondary school students: Profiles and antecedents. *Journal of Numerical Cognition, 5*(3), 283-300. https://doi.org/10.5964/jnc.v5i3.201

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1). 67-72. https://doi.org/10.4103/aca.ACA_157_18

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159-168. https://doi.org/10.1207/s15327574ijt0502_4

Niemivirta, M., Pulkka, A., Tapola, A., & Tuominen, H. (2019). Achievement goal orientations: a person-oriented approach. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge Handbook of Motivation and Learning* (pp. 566-616). Cambridge University Press. https://doi.org/10.1017/9781316823279.025

OECD [Organisation for Economic Co-operation and Development]. (2019a). *Programme for International Student Assessment (PISA) Results from PISA 2018: Finland Country Note.* https://www.oecd.org/pisa/publications/PISA2018_CN_FIN.pdf

OECD [Organisation for Economic Co-operation and Development]. (2019b). *Programme for International Student Assessment (PISA) Results from PISA 2018: United States Country Note.* https://www.oecd.org/pisa/publications/PISA2018_CN_USA.pdf

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. Routledge.

Rinne, L. F., Ye, A., & Jordan, N. C. (2017). Development of fraction comparison strategies: A latent transition analysis. *Developmental Psychology, 53*(4), 713-730. https://doi.org/10.1037/dev0000275

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well-taught" mathematics courses. *Educational Psychologist, 23*(2), 146-166. https://doi.org/10.1207/s15326985ep2302_5

Selter, C. (2009) Creativity, flexibility, adaptivity, and strategy use in mathematics. *ZDM – Mathematics Education, 41*(5), 619-625. https://doi.org/10.1007/s11858-009-0203-7

Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental Psychology, 49*(10), 1994-2004. https://doi.org/10.1037/a0031200

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*(4), 273-296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Sievert, H., van den Ham, A., Niedermeyer, I., & Heinze, A. (2019). Effects of mathematics textbooks on the development of primary school children's adaptive expertise in arithmetic. *Learning and Individual Differences, 74*, Article 101716. https://doi.org/10.1016/j.lindif.2019.02.006

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Threlfall, J. (2009). Strategies and flexibility in mental calculation. *ZDM – Mathematics Education, 41*(5), 541-555. https://doi.org/10.1007/s11858-009-0195-3

TIMSS & PIRLS International Study Center. (2021). *Countries' Mathematics and Science Achievement: Mathematics Grade 8.* https://timss2019.org/reports/achievement/#math-8

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2006). The development of children's adaptive expertise in the number domain 20 to 100. *Cognition and Instruction, 24*(4), 439-465. https://doi.org/10.1207/s1532690xci2404_2

Vamvakoussi, X., & Vosniadou, S. (2010). How many decimals are there between two fractions? Aspects of secondary school students' understanding of rational numbers and their notation. *Cognition and Instruction, 28*(2), 181-209. https://doi.org/10.1080/07370001003676603

Van Hoof, J., Vamvakoussi, X., Van Dooren, W., & Verschaffel, L. (2017). The transition from natural to rational number knowledge. In D. C. Geary, D. B. Berch, R. J. Ochsendorf & K. M. Koepke (Eds.), *Acquisition of Complex Arithmetic Skills and Higher-Order Mathematics Concepts* (pp. 101-123). Academic Press. https://doi.org/10.1016/B978-0-12-805086-6.00005-9

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982) A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17*(3), 371-388. https://doi.org/10.1207/s15327906mbr1703_5

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2009). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education. *European Journal of Psychology of Education, 24*(3), 335-359. https://doi.org/10.1007/BF03174765

Voyer, D., Nolan, C., & Voyer, S. (2000). The relation between experience and spatial performance in men and women. *Sex Roles, 43*(11/12), 891-915.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250-270.

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219-246. https://doi.org/10.1177%2F0095798418771807

Wertheimer, M. (2020). *Productive Thinking* (V. Sarris, Ed.). Birkhäuser. (Original work published 1945)

Zajic, Al. (2019, December 28). Introduction to AIC – Akaike Information Criterion. *Beyond Data Science*. https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced