Promoting the Development of an Integrated Numerical Representation through the Coordination of Physical Materials

Jonathan Vitale

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Promoting the Development of an Integrated Numerical Representation through the Coordination of Physical Materials

Jonathan Vitale

How do children use physical and virtual tools to develop new numerical knowledge? While concrete instructional materials may support the delivery of novel information to learners, they may also over-simplify the task, unintentionally reducing learners' performance in recall and transfer tasks. This reduction in testing performance may be mitigated by embedding physical incongruencies in the design of instructional materials. The effort of resolving this incongruency can foster a richer understanding of the underlying concept. In two experiments children were trained on a computerized number line estimation task, with a novel scale (0-180), and then asked to perform a series of posttest number line estimation tasks that varied spatial features of the training number line. In experiment 1, during training with feedback, children either received a ruler depicting endpoint and quartile magnitudes (i.e., 0, 45, 90, 135, 180) that physically matched the on-screen number line (*congruent ruler)*, a proportionally-similar ruler scaled 33% larger than the on-screen number line (*incongruent ruler)*, or no ruler. Children were trained to criterion before proceeding to posttest. Results indicated that while children who used the *congruent ruler* performed well during training, their performance at posttest was less accurate than the other two conditions. On the other hand, by increasing the difficulty of the learning task, while providing relevant landmark information, children in the *incongruent ruler* condition produced the highest accuracy at posttest. In experiment 2, controlling for learning task duration, the *incongruent ruler* and *congruent ruler* conditions were compared directly. Posttest results confirmed an advantage for children in the more complex, *incongruent ruler* condition. These results are interpreted to suggest that landmarks representations are an important and accessible means of developing a mature numerical representation of the number line. Furthermore, the results confirm that desirable difficulties are an essential component

of the learning process. Potential implications for the design of learning activities that balance instructional support with conceptual challenge are discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Dr. John Black. I truly appreciate the opportunity I have had to pursue my interests and engage exciting research. Thank you Dr. Herbert Ginsburg and Dr. Matthew Johnson for helping me organize and refine my ideas from proposal hearing to defense. In particular, thank you Dr. Ginsburg for pushing me to think outside of my own quantitative box, I hope to apply these ideas for years to come. Thank you to Dr. Sandra Okita and Dr. Koleen McCrink for your valuable insights during my dissertation hearing. Finally, thank you Dr. Robert Siegler for all of the advice and encouragement throughout this process. I hope that my work reflects the high standards that you have set in this line of research.

I would also like to thank all of my fellow students that helped me considerably over the course of my graduate education. Thank you to Eric Carson, Tim Chang, and Archana Vaidyanathan for helping me collect data. Thank you to Lance Vikaros, Genevieve Hartman, Jessica Hammer, Cameron Fadjo, and Jamie Krenn for the advice, constructive criticism, and general food-for-thought. Thank you to Michael Swart for all of the above and all of the additional work you have done for this project.

I chose to pursue this degree so that I could make a positive impact in children's lives. While I cannot be sure that I have accomplished this goal, yet, the children of New York City have made an amazing impact on my life. Thank you to all of the children from the High School for Health Careers and Sciences that serve as a continuous source of grounding for my ideas. Thank you to all of my lifelong teacher friends. Thank you to the students and staff of PS 115, CS 154, and PS 161 who have been an integral part of this process. In particular, to Grace and Jackie from PS 115 and Tom from CS 154, I could not have done this without you.

Finally, I would like to thank my family. My parents have supported my education from pre-school to Ph.D. I owe my *focus* to you. Lastly, I would like to thank my wife, Lisa Caswell, for being caring, encouraging, patient, and all around spectacular. I can finally come home for dinner!

DEDICATION

I dedicate this dissertation to my daughter, Rosalie Anna Vitale. I know that you will develop into an amazing, independent, inquisitive woman.

1. Introduction

External representations of mathematical concepts – such as an equation, a number line, or a graph – not only serve to make difficult computations simpler, but support and guide the development of our internal, conceptual representations. This interactive relationship between internal and external representation reflects Piaget's (1952) notion of adaptation, in that knowledge is either assimilated into our existing cognitive structure or our cognitive structure is transformed to accommodate new knowledge. While accommodation is often necessary when a conceptual representation lacks sufficient power to address new problems, assimilation is the primary and often preferred form of adaptation for children (Ohlsson, 2009; Piaget, 1962). Therefore, learners are likely to interpret novel or modified instructional tools in terms of known concepts, intuitive strategies, and contextual elements – perhaps at the cost of deep, conceptual change. Instruction is intended to promote adaptation to formal concepts.

Yet, Piaget (1952) views the developmental process as auto-regulative, such that learning – particularly in the form of large-scale conceptual change – is the result of a self-determined, rigorous intellectual struggle with the material at hand. Attempts to accelerate development through direct instruction may only lead to superficial knowledge. Alternatively, a loosely-structured learning environment may provide opportunities for children to develop deeper knowledge through a series of self-motivated inquiries. While this approach retains persistent popularity, a number of researchers point to mixed or negative results of discovery-based learning (Anderson, Reder, Simon, Ericsson, & Glaser, 1998; Kirschner, Sweller, & Clark, 2006; Mayer, 2004). Furthermore, a large, diverse body of research shows that conceptual development can be promoted through guided instruction – which may even affect performance in Piagetian stage-determining tasks, such as magnitude conservation (Brainerd, 1972; Gelman, 1969; Wallach & Sprott, 1964).

While instruction can spur cognitive development, it can also limit or even prevent learning. In his work on “desirable difficulties”, Robert Bjork and colleagues (e.g. Bjork & Bjork, 2011; Bjork & Linn, 2006; Bjork, 1994) argue that unchallenging instruction tasks, which produce successful performance during learning, may limit long-term retention and transfer. In the domain of mathematical instruction, if a difficult new concept is introduced in a highly-structured environment, such that the problem may be solved solely

in terms of intuitive or well-rehearsed strategies, conceptual change is unlikely (Martin & Schwartz, 2005; Schwartz, Varma, & Martin, 2008). As Schwartz, Varma, and Martin (2008) argue, transfer often requires the creative coordination and integration of multiple concepts and strategies. If the learning task solves the coordination problem itself, the child misses the opportunity to develop a deeper understanding of the concept. Rather, by introducing materials that embed reasonable impediments to coordination – "incongruencies" – to promote higher-level, reflective processes, learners are more likely to develop flexible, robust concepts.

In essence this assertion entails that there needs to be an appropriate balance between structure and learner independence. While the learning activity needs to be carefully designed to ensure that appropriate challenges emerge, only the learner himself or herself can meet these challenges. In the remainder of this chapter I will detail this need for balance by describing studies that (1) establish the need for limited guidance in instruction, and (2) demonstrate the types of desirable difficulties that may enhance conceptual representation. In the chapter that follows I will apply these ideas to the specific domain of numerical cognitive development.

1.1 Guiding attention for instruction

While Piaget (1970), Bruner (1966), and other constructivist researchers (e.g. Papert, 1980) argue that the child plays a primary, productive role in his or her own development, specific instructional interpretations of these arguments show mixed results (Mayer, 2004). Specifically, research on discovery learning fails to show significant achievement gains in areas of problem-solving (Gagné & Brown, 1961; Kittel, 1957; Shulman & Keisler, 1966), conservation strategies (Brainerd, 1972; Gelman, 1969; Wallach & Sprott, 1964), and computer programming with LOGO (Fay & Mayer, 1994; Lee & Thompson, 1997; Pea & Kurland, 1984).

As Kirschner et al. (2006) explain, one challenge with discovery-based learning is that the task demands may overwhelm working memory. Too much cognitive load on a problem-solving task can lead to poor performance and little learning (Sweller, 1988). Furthermore, given the large "search space" of many ill-defined problem solving or learning tasks, one may simply miss the pertinent elements of the task. However, as Blanchette and Dunbar (2000) explain, failure to transfer in these situations does not

imply that the learners did not (or could not) learn the material, but that they simply did not to generate the specific representation that the task designer had in mind. From this perspective, some level of instructional guidance is necessary to put learners on the right path.

For example, in standard problem solving tasks, such as those presented by Gick and Holyoak (1980, 1983) – utilizing Duncker's (1945) problems – adults are often unable to transfer solution strategies between two superficially dissimilar, but analogical problems. Specifically, adults who read a story about a military operation, utilizing a distributed application of force, were asked to solve an analogous problem with a characteristically dissimilar cover story (i.e., a tumor surrounded by healthy skin). Unless prompted to reflect upon the original story, or asked to compare and contrast story analogs, a majority of participants were unable to correctly solve the tumor radiation problem.

In a more recent eye-tracking approach to the problem Grant and Spivey (2003) found that just prior to solving an illustrated version of the radiation problem a surprising number of participants focused on the perimeter ("skin") of the healthy tissue surrounding the tumor. In a follow-up study Grant and Spivey used animation to highlight the skin; thus drawing attention to the critical feature. In comparison to the (approximately) one-third of participants who solved the problem with a static image, two-thirds of the participants solved the problem in the animated condition. The large shift in success, prompted by a minimal highlighting manipulation, suggests that problem solving success is, in some cases, an arbitrary function of chance – i.e., happening to focus attention on an important feature.

Similarly, in instructional settings, without appropriate guidance, children may overlook critical structural features of the content – particularly if that task incorporates highly realistic or complex elements (Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler, 2008; Son & Goldstone, 2009). In particular, younger children may place greater attention on superficial aspects of materials (Quinn & Eimas, 1997). This shift in attention from superficial features to abstract concepts is not only a general developmental trend, but can occur between novices and experts in a field. For example, while novice physics students sort physics problems based on superficial, perceptual characteristics, physics experts sort problems based on abstract, structural features (Chi, Feltovich, & Glaser, 1981).

While abstraction is a reasonable goal of education, instruction that is based in highly symbolic, abstract materials may appear highly disconnected from real world activity, resulting in low transfer to

authentic contexts and disengaged students (Lave, 1988). Furthermore, while knowledge may be abstract, in the sense of being generally applicable rather than contextually-bound, the actual components of internal representations may be perceptual in nature (Barsalou, 2003). Given this close relationship between perception and conception (Goldstone & Barsalou, 1998), promoted in the field of "embodied" or "grounded" cognition (Barsalou, 2008; Clark, 1999; Glenberg, 1999; Goldstone, Landy, & Son, 2010; Thelen, 2000), the structural features of a concept may be conveyed best by training perceptual or motor systems (Black, Segal, Vitale, & Fadjo, 2012; Gibson, 1969; Goldstone & Son, 2008; Goldstone et al., 2010). This may be particularly true in inherently spatial domains, like mathematics (Lakoff & Núñez, 2000).

For example, Goldstone and colleagues (Goldstone & Wilensky, 2008; Goldstone & Son, 2005; Son & Goldstone, 2009) applied a technique called "concreteness fading" to teach complex systems, such that learners were exposed to progressively schematic instantiations of a simulated complex system. While this "abstraction" remains perceptual in nature, salient but irrelevant features of the simulation are discarded. Learners' internal representations are hypothesized to follow this shift towards greater schematization and generality.

Although concreteness fading is effective, it requires a large degree of control over the learning materials. In many cases, the activity designer may prefer to focus on adding attention-guiding visuals or instructions that overlay the task, rather than change the structure of the task itself. One such technique, called "signaling" (Mautone & Mayer, 2001, 2007), increases the perceptual salience of some feature of the display – which, in turn, should increase the conceptual salience of the feature. Signaling may be particularly useful in applications where multiple features compete for attention. For example, in text, signals may include the manipulation of fonts, type size, headings, or the inclusion of additional "pointer" words (Loman & Mayer, 1983; Lorch & Lorch, 1995, 1996; Lorch, Lorch, & Inman, 1993; Mautone & Mayer, 2001; Meyer, 1975). In a spatial context, the Grant and Spivey (2003) example, described above, demonstrates how a visual signal can draw attention to physical structures.

Likewise, Mautone and Mayer (2007) promote the incorporation of graphical organizers, which place a layer of structure over the instructional material. For example, Stull, Hegarty, and Mayer (1999), implement this type of organizer in the study of anatomy learning in a virtual reality environment. In this

experiment, medical students were asked to rotate virtual 3D anatomical structures to a target orientation with or without orientation references – i.e., segments drawn along orthogonal, but arbitrary, axes of the structure. Expectedly, with difficult rotations, participants with orientation references rotated their figures more accurately and efficiently. Furthermore, low spatial-ability participants performed better on a test of anatomical knowledge – at nearly the same level as high spatial-ability participants – with access to these references in the rotation task than without.

In educational settings the use of "spatial tools" (Mix, 2009), are often designed to highlight or depict the critical structural features of a concept. For example, base-10 blocks are a common means to demonstrate the relationship between place values in standard elementary curricula (Fuson & Briars, 1990). Likewise, in the study reported here, children used a common spatial tool (i.e., a ruler) to guide their behavior, and potentially shape their representation of the target concept.

1.2. Desirable difficulties in instructional activities

Clearly, providing learners with spatial tools (also known as concrete manipulatives, Sarama & Clements, 2009) to guide perception and prompt specific cognitive strategies is an important element of effective instruction. This explains manipulatives' central – and in some cases unrealistic (Ball, 1992) – role in constructivist curriculum (McNeil & Uttal, 2009). Yet, research on concrete manipulatives yields mixed results, at best (Brown, McNeil, & Glenberg, 2009; Fennema, 1972; Fuson & Briars, 1990; Hiebert et al., 1996; Martin, Lukong, & Reaves, 2007; Meira, 1998; Moyer, 2001; Uttal, Doherty, Newland, Hand, & Deloache, 2009).

While the reasons why manipulatives vary in effectiveness are diverse (see McNeil & Jarvin, 2007), it may be the case that, in contrast to the examples given above, some manipulatives provide too much behavior-leading structure (Martin & Schwartz, 2005). Specifically, by reducing the difficulty of the task, excessive structure may interfere with critical cognitive processes that support the encoding of robust memory representations of the target concept (Bjork, 1994). Thus, from the perspective of "desirable difficulties" (Bjork & Linn, 2006; Bjork & Bjork, 2011; Bjork, 2006; Bjork & Linn, 1999; Linn, & Bjork, 2007), some degree of failure is necessary in learning environments.

While much of the focus on desirable difficulties is on how the process of forgetting (and later re-encoding) produces stronger memories than learning events that do not support forgetting (Anderson, Bjork, & Bjork, 2000; Anderson, Neely, Bjork, & Bjork, 1996; Bjork, 1994; Hays, Kornell, & Bjork, 2010; Kornell & Bjork, 2009; Storm, Angello, & Bjork, 2011), learning manipulations that foster more complex, deeper cognitive processing – e.g. generative processes (Richland et al., 2007; Schwamborn, Mayer, Thillmann, Leopold, & Leutner, 2010) – may fall under this heading (Bjork & Linn, 2006). Specifically, there are several pertinent examples in which some inconsistency or ambiguity is explicitly incorporated into the learning materials, which promote greater higher-level thinking than simpler materials.

For example, in a study of reading comprehension, Mannes and Kintsch (1987) tested the effects of outlines on participants' recall and understanding of a text. In one case the outline was consistent with the layout of the chapter, and in the other case the outline was inconsistent with the layout of the chapter. As might be expected, participants who read the consistent outline were more likely to successfully recall verbatim facts from the text; however, participants who read the inconsistent outline were more likely to successfully answer inference questions that relied on a deeper understanding of the material. The authors suggest that because the consistent outline provides an intact "macrostructure" for the text, participants devoted more cognitive resources to learning specific propositions. On the other hand, participants with an inconsistent outlined devoted more cognitive resources to constructing the appropriate macrostructure, which maintains the gist of the text.

Similarly, in the domain of mathematics, Martin and Schwartz (2005) asked students to engage in a number of fraction exercises with either a set of identical rectangular ("tile") pieces, or a set of diverse circular section ("pie") pieces. In the case of the pie, both the numerator and denominator were inherent properties of the shape of each piece (e.g. a semicircle is always ½). On the other hand, each rectangular piece's value could only be determined by its relationship to the other pieces. For example, summing ½ and ¼ with pie pieces is simply a matter of combining the correct pieces and counting the number of quarters filled. On the other hand, with tile pieces, one must first determine how to represent ¼ and ½ with a set of equivalent rectangles. The additional problem of correctly representing the task could be interpreted as a limitation of the tile pieces; however, the children who learned with the tile pieces were significantly more likely to transfer learned strategies to novel problems.

Finally, in a recent study, Byrge and Goldstone (2011) tested the effects of learning a mechanical system on transfer to a population dynamics problem. In a computer simulation-based learning task, adult participants manipulated an animated fan to apply a force on an oscillating ball. Participants manipulated a slider to activate the fan. While in the "compatible" condition the direction of the slider was aligned with the direction of the fan's force, in the "incompatible" condition the direction of the slider opposed the direction of the fan's force. In the transfer task, participants could choose to apply "media" to affect the natural ebb and flow of the population within a city. The same general strategy – i.e. to apply a "force" during upward swings of the system – maximized the amplitude of each system.

While participants in both condition found the transfer task to be initially challenging, only those in the incompatible condition, on average, improved. The authors suggest that by encouraging an incongruent gesture participants were forced to mentally "tease-apart" the relationship between the action of the slider and the action of the fan, resulting in a more flexible representation. In the compatible condition, on the other hand, the two types of actions blurred together to produce a shallower representation and a functional fixedness in the subject. Alternatively, compatibility may have allowed the participants to ignore their own actions altogether – relegating the gesture to a subconscious routine – and instead focus on the visual depiction exclusively. While making a physical action intuitive is a common goal of user interfaces, in this case critical attention to one's own action was a necessary element of learning.

In all three examples here, the seemingly "well-designed" materials, in which the structural organization of the target concept was conveyed through an intuitive model, lead to decreased transfer. In the case of visualizations, the reliable, intuitive design conveyed "deceptive clarity" (Linn, Chiu, Zhang, McElhaney, & McElhaney, 2010), which fostered overconfidence and discouraged learners from engaging in deeper reflection of the material. On the other hand, materials that incorporated a conceptual inconsistency or ambiguity promoted transfer. In this case, the challenge of coordinating disparate materials (e.g. summary and text, graphical model and symbolic representation), prompted the type of deeper cognitive processes, or "knowledge integration" (Clark & Linn, 2003), that result in successful transfer.

Yet, this notion that transfer may be promoted by encouraging learners to reconcile seemingly incompatible representations (both internal and external), deviates from the more common “structure-mapping” account (Gentner, 1983, 2010), in which transfer occurs through a direct mapping of a known concept to a structurally-similar, but novel conceptual space. In their account of “dynamic transfer” Schwartz et al. (2008), argue that this form of concept-building, in which the resulting representation is distinct from the source concepts, is necessary for innovation. This idea echoes Fauconnier’s concept of “conceptual blends” (e.g. Fauconnier & Turner, 1998; Fauconnier, 1994; Fauconnier & Turner, 2002), in which two or more source representations are integrated, non-linearly, into a novel representation. The notion of conceptual blending and integration is a critical feature of Lakoff & Núñez' (2000) account of mathematical development, which I discuss in the subsequent chapter.

2. Numerical estimation

While the studies described above show the effects of instruction on advanced concepts, given the persistent relationship between early mathematical school readiness and later achievement (Duncan et al., 2007), there is also a need to focus on foundational concepts. In recent years a number of researcher have demonstrated a direct link between the "number sense" Dehaene (1997) and mathematical achievement (Booth & Siegler, 2006; Halberda, Mazzocco, & Feigenson, 2008; Holloway & Ansari, 2009). According to a number of researchers, mathematical knowledge and ability is embodied in core perceptual/spatial processes (Feigenson, Dehaene, & Spelke, 2004; Lakoff & Núñez, 2000). For example, Halberda et al. (2008) found that children's ability to discriminate between two dot displays by quantity (i.e., choose the set with "more" dots), was predictive of achievement scores across a range of years.

However, in other studies, such as that conducted by Holloway and Ansari (2009), individual differences appeared to be founded less on raw perceptual capabilities than on how these capabilities were integrated with symbolic representations. Case and Griffin (1990) found that low-SES children, at greatest risk of failure in mathematics, often applied "non-adaptive" strategies to numerical computations due to their poorly developed sense of number. Additionally, Griffin, Case, and Siegler (1994) found that these children's knowledge representations could be greatly enhanced by focusing on activities that integrate number sense-building and computation. Therefore targeting the development of central structures (e.g. the number line) in interventions is likely to result in dramatic effects (Case et al., 1996; Griffin, Case, & Siegler, 1994).

Designing appropriate instructional interventions requires a clear understanding of how intuitive forms of knowledge develop and how they are applied, or misapplied, to common mathematical tasks. In recent years a large body of research has emerged to address issues of number sense and its relationship to computation.

2.1. Initial development of number sense

Perhaps surprisingly, one's sense of magnitude emerges early, prior to the onset of language (Feigenson, Dehaene, & Spelke, 2004). Specifically, infants are capable of nonsymbolic numerosity

discrimination between sets of objects that have a large proportional difference. For example, six-month-olds can discriminate between 8 and 16 dots, but not 8 and 12 (Xu & Spelke, 2000). The capability to discriminate between stimuli with smaller magnitude ratios increases with older infants; e.g. nine-month-olds can discriminate between 8 and 12 sounds (Lipton & Spelke, 2003).

Beyond perceptual discrimination of non-symbolic stimuli, the effects of magnitude extend to common symbolic representations – effecting both duration and accuracy across a range of numerical tasks (Dehaene, 1997). Brain imaging studies suggest that a region within the parietal lobe – the intraparietal sulcus (IPS) – supports both symbolic and nonsymbolic magnitude or quantity comparisons (Fias, Lammertyn, Reynvoet, Dupont, & Orban, 2003; Pinel, Piazza, Bihan, & Dehaene, 2004). This area of the parietal cortex is often associated with spatial processes and shows frequent activation during a variety of numerical tasks varying numeric representation (e.g. Arabic numerals, dot arrays), sensory modality, and motor response type (Dehaene, Piazza, Pinel, & Cohen, 2003; Hubbard, Piazza, Pinel, & Dehaene, 2005).

This association between space and number is interpreted by some as evidence of a “mental number line”, encoded in the IPS, that drives numerical processing (Dehaene & Cohen, 1995; Dehaene, 1997; Dehaene, Bossini, & Giraux, 1993). Behaviorally, the mental number line is perhaps most clearly manifested in the SNARC effect by biasing ones motion towards the left when small numbers are presented and towards the right side when large numbers are presented in any numerical task (Dehaene et al., 1993). Corroborating neuroanatomical evidence suggests that a relationship exists between the physical layout of number-sensitive cells in the IPS and perceived numerical magnitude (Hubbard, Piazza, Pinel, & Stanislas Dehaene, 2005).

Yet, while the mental number line provides a foundation for numerical calculation, its structure does not conform to the normative, linear distribution of whole numbers. Rather, reflecting Fechner’s Law, one model of the mental number line organizes the relationship between actual and perceived magnitude according to a logarithmic function (Piazza et al., 2004). This logarithmic model entails that smaller numbers are granted more representational space than larger numbers – making differences between small numbers (e.g. 1 and 2) more salient than differences between larger numbers (e.g. 13 and 14). While logarithmic effects of magnitude may only emerge with adults in rapid tasks or in duration

analyses (Viarouge, Hubbard, Dehaene, & Sackur, 2010), a logarithmic function typically provides a better fit for five-year-old children's overall judgment of magnitude on a number line estimation task than a linear function (Siegler & Opfer, 2003).

While a logarithmically-compressed mental number line suits a functional purpose – i.e. more precision with smaller numbers, which permeate our cultural environment in greater abundance than large numbers (Dehaene & Mehler, 1992) – the application of a logarithmic representation is inappropriate for most mathematical tasks, including number line estimation. Adults' number line estimates are typically distributed linearly across common numerical scales (Siegler, Thompson, & Opfer, 2009). Yet, while the development of early, pre-verbal number sense is a continuous developmental process, the progression from logarithmic to linear representation is marked by abrupt, qualitative changes as children learn to apply their linear representation to new scales (Opfer & Siegler, 2007; Siegler & Booth, 2004; Siegler et al., 2009). Given the relationship between number line estimation and mathematical achievement (Booth & Siegler, 2006), instructional effects on numerical representations is a promising target of research.

2.2. Promoting mature representations of the number line

In studying preschool children's ability to map numbers to an analog representation (number line) Le Corre & Carey (2007) found that knowledge of counting principles (CP-knowledge) – such as cardinality, stable-order, and one-to-one (Gelman & Gallistel, 1978) – was a necessary but insufficient condition for successful analog mapping. Rather, the development from CP-knowledge to analog mapping required additional awareness of how numerical principles interface with analog, spatial tasks.

Understanding how to apply a count-based understanding of number to analog mapping may be promoted with activities, such as board games, in which a child counts the motion of an object along a linear path. Several interventions applying this approach, within the 1-10 scale, show that a normative, linear representation may develop rapidly with preschool children (Ramani & Siegler, 2008; Siegler & Ramani, 2008; Siegler & Ramani, 2009). Therefore, by transferring knowledge of counting principles to an analog task, children extend their concept of number and how it may be applied.

Yet, while this approach may be sufficient at small scales, at a much larger scale, e.g. 0 – 100, explicit counting behaviors become less relevant. Rather, children must rely on the core perceptual capabilities of the approximate number system. As such, this often results in task performance that may be described in terms of logarithmic relationship between estimated and actual magnitude (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003). Likewise, when asked to hold small numbers in working memory, adult performance in a numerical task (number line bisection) tends to show evidence of a logarithmic representation (Lourenco & Longo, 2009).

Adults can overcome this intuitive bias towards the logarithmic scale through explicit attention to known structural features of a given numerical scale. For example, Opfer and Siegler (2007) found that by simply providing feedback on estimation trials, at magnitudes where the discrepancy between the logarithmic and linear representations is the greatest, children underwent a rapid, holistic shift from a logarithmic representation to a linear representation on the given numerical scale. In a subsequent study Thompson and Opfer (2010), applied the technique of "progressive alignment" to encourage transfer from a presumed linear representation at the 0-100 scale to the 0-1000, 0-10,000, and 0-100,000 scales, with participants of various ages. By learning to recognize the structural similarities between these scales (i.e., similarity of leading digits), participants successfully mapped a known representation to a new conceptual space.

Yet, it is difficult to conceive how this approach might be applied to a numerical scale that differs in its leading digits, e.g. 0 – 180. Rather, adults who successfully engage in an estimation task with a novel numerical scale or spatial layout must be applying some numerically- and spatially-generic strategy. As evidence, in their original report of the logarithmic-to-linear shift, Siegler an Opfer (2003) found significantly less variance (greater precision) in the estimates of magnitudes near quartile points of the numerical scale (i.e., 25, 50, and 75 in a 0 – 100 scale). This finding suggests that adults either implicitly (perceptually) or explicitly (strategically) divide the number line into sections while estimating (i.e., "divide-and-conquer"). In an alternative model, utilizing a cyclical power function rather than a linear or logarithmic function, these same quartile points ("reference points") are hypothetical locations for breaks between cycles (Barth & Paladino, 2011; although, see Opfer, Siegler, & Young, 2011).

Regardless of the underlying mathematical model, landmark values may play an important role in the development of mature representations. Specifically, evidence of abrupt shifts in representation (Opfer & Siegler, 2007) and adult bias towards logarithmic representations (Lourenco & Longo, 2009), suggest that a linear representation does not completely and permanently replace a logarithmic representation. Rather, one possibility is that adults generate an online linear representation for the task at hand by integrating both their intuitive sense of numerical magnitude and explicit knowledge of structural features – such as landmarks. From this perspective, the mature representation may be described as a conceptual blend between these various source representations (Lakoff & Núñez, 2000). As such, promoting a landmark-based representation of the number line should lead to linear estimation.

2.3. Developing landmark-based representation: Attempt 1

The role that landmark-based representations of the number line play in the natural trajectory from immature (logarithmic) to mature (linear) representations is a potentially important focus of research. If it is the case that landmark-based representations are either necessary or sufficient for a linear representation, then by promoting the internalization of landmark spatial-numerical associations (e.g. 50-as-midpoint) a child's estimates may become, permanently, more like those of adults. Furthermore, because of the generality of the midpoint concept, a landmark-based representation might provide a stronger basis for transfer to spatially-dissimilar tasks.

To address these hypotheses, in a previous study, I developed a numerical game in which children estimated a distance along the number line within a lake fishing narrative (Vitale, Siegler, & Black, 2011). This game was applied across four conditions, with two factors: linearity of display (linear or circular number line) and presence of landmarks (endpoint-only or quartile). In the case of the quartile landmarks, the hatch marks along the number line were persistantly displayed through game play along with the two closest numerical values to the cursor's position (see Fig. 1).

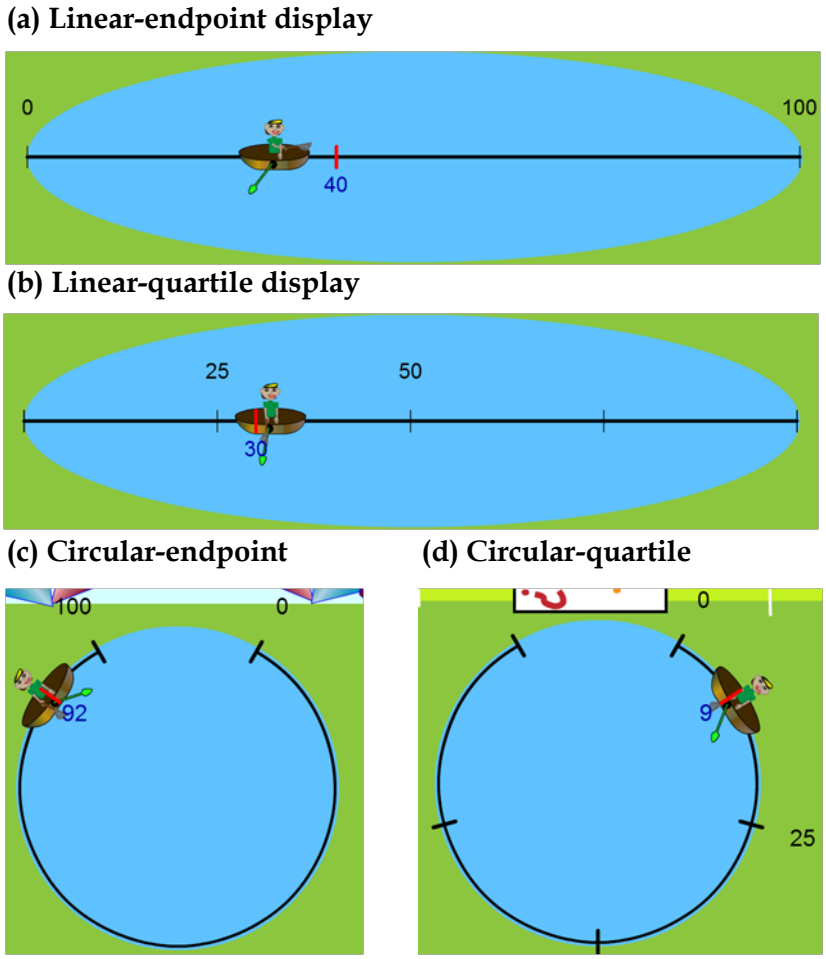


Figure. 1: Number line game. Each condition of the experiment is displayed during the feedback sequence, between trials. During trials, background features are hidden to avoid visual distraction. Images are cropped.

To test the effect of training on linear estimation, four groups of kindergarten students were trained for three 20 minute sessions on a single display condition. Children also received a number line estimation pretest and posttest, with only 0 and 100 printed as reference points. As expected, during training those children with access to quartile information (hatch marks and associated values) were successful for a greater percentage of trials than those children with only endpoint information. Also, when analyzed in terms of coherence of estimates to a linear distribution ("linearity"), children with quartile information also outperformed those children with only endpoint information, demonstrating that children can effectively apply landmark information to either a linear or nonlinear display.

However, at posttest this advantage for the quartile displays vanished. In an analysis of the median estimates verses actual magnitude, only children's estimates in the linear-endpoint condition

demonstrated the expected shift from a logarithmic to a linear representation. In other measures of accuracy, including mean errors, no difference between landmark and non-landmark conditions arose, while large differences between the linear and circular conditions were clearly evident.

This evidence suggests that providing persistant landmark information during training is, at best, inconsequential for the development of strong, flexible internal representations. At worst, presenting this additional information interferes with learning. However, dismissing landmark displays as a potentially valuable tool is premature. Their effectiveness during game play suggests that the children either were unable to recall the spatial position or numerical value of the landmarks, or that they had difficulty coordinating a strategy with recalled (imagined) landmarks at posttest.

Bjork (1994), describes how once a memory is activated repetition does little to strengthen its encoding (i.e., minimal benefits of “overlearning”). As such, the persistant structure of landmarks provided little opportunity to strengthen the memory through forgetting and re-encoding. On the other hand, with the endpoint-only conditions, forgetting between trials was more likely and may have lead to strengthening the memory of this spatial-numerical associations.

Yet, even though children in the endpoint-only, linear condition produced, by some measures, more mature representation at posttest, it cannot be assumed that this knowledge could be applied flexibly. Given that the children in the endpoint-only, circular display condition could not transfer to a spatially-different number line, it is also likely that children in the linear display condition would have encountered a simliar problem transfering to a spatially discrepent display (e.g. a circular number line). Specifically, children may have developed associations between specific numbers (or ranges of numbers) and absolute position on the number line. For example, a child could have arbitrarily recalled where he or she had seen the number “37” on screen in a recent trial. While this would likely be an effective strategy for the unchanging learning task or a highly similar posttest, the knowledge would be inapplicable to a new numerical scale or spatially dissimilar number line.

2.4. Incorporating a desirable difficulty

The possibility that children in the endpoint-only condition, even when succesful, may have constructed a highly contextually-bound representation of numerical magnitude is an important rationale

for promoting a spatially-generic, landmark-based representation. On the other hand, providing landmark information, in the previous manipulation, substantially increased learning task performance at the cost of long-term retention. The “deceptive clarity” (Linn et al., 2010) of this landmark visualization discouraged the active, generative role that children must play in integrating an intuitive and landmark-based representation of number. To make the integration process explicit, instead of embedding landmarks in the number line itself, in the following experiment children were given a second physical representation of the number line – i.e., a ruler – that acted as a physical manifestation of an independent concept of number.

Yet, if a child placed the ruler directly against the number line, the resulting, physically integrated representation would be equivalent to the digitally integrated display described previously. Once again, the simplicity of the task would likely reduce learning outcomes. Instead, by providing a ruler that was spatially *incongruent* to the target number line the child’s attention is re-directed to the process of coordinating materials. In this case, a slighty longer ruler retained the proportional relationships between landmark values, while preventing direct physical mapping of landmark values to the on-screen display. This *incongruent ruler* thus represents a compromise between minimal guidance – which may lead to development of a contextually-bound representation – and over-structuring – which may lead to low retention.

3. Experiment 1

In this study I target the development of a landmark-based representation of the number line. As described above, addressing this concept effectively requires a suitable balance between external structure and independent learner action. In the case where there is not enough structure – i.e., where the child is not provided with any physical representation of landmarks (*no ruler*) – we should expect low efficiency in learning, and attention to irrelevant features (i.e., low transfer). In the case of too much structure – i.e., where the child is provided a physical representation of landmarks (a ruler) that is congruent in physical length to, and thereby easy to integrate with, the on-screen number line – we should expect strong performance in the learning task, but poor performance once the scaffolds have been removed for the posttest. In a third case, by incorporating an impediment to the coordination of landmarks with the number line – i.e., by making the physical representation of landmarks (the ruler) spatially incongruent with the on-screen, target number line – we should expect strong performance in transfer activities, at some cost to efficiency in learning (vs. the congruent ruler). In this study children were tested on an equivalent number line to the testing display, and three number lines that altered some spatial feature(s) of the original display.

As in many educationally-relevant experiments, choosing the most appropriate experimental parameter to equate learning experience between conditions was an important consideration. In this study, there is an a priori assumption of differences in the inherent difficulty of treatments. To measure transfer of knowledge attained through mastery of the particular learning environment, I chose to train children to criterion (see Chariker, Naaz, & Pani, 2011, for a recent study with a similar approach). While this choice entailed that participants could have engaged the task for differing number of trials, it ensured that no child would proceed to testing without having a sufficient knowledge of the learning environment. Differences in measures of the training experience also provided suitable variables to test hypotheses about learning efficiency.

Specifically, the following hypotheses describe the predicted relationship between conditions:

(1) Children provided with a tool that provided the most intuitive structure for the task, i.e., the *congruent ruler,* should demonstrate the greatest efficiency in the learning task but the least successful transfer to the posttest, compared to the other two conditions.

(2) Children provided limited, but conceptually challenging guidance during the learning task, i.e., the *incongruent ruler,* should demonstrate both higher efficiency in the learning task as well as greater accuracy at posttest than children with minimal guidance, i.e., *no ruler.*

(3) Finally, while children in the *incongruent ruler* condition are expected to outperform children in the *no ruler* condition at posttest in general, this effect should only emerge after the first subtest. Specifically, because the number line used in first subtest directly matches the spatial and numerical layout of the training display, children trained to criterion from both conditions should do equally as well. Beyond the first subtest, the inflexibility of the representation afforded by the *no ruler* condition should substantially weaken their performance relative to the spatially-generic representation afforded by the *incongruent ruler* condition.

3.1. Method

3.1.1. Participants

Participants included a total of 80 second, third, and fourth grade students. Children were gathered from two organizations in the New York City area whose services include a daily afterschool program that extends into a summer day camp during July and August. Seventy-five children were recruited from one afterschool/summer school program hosted at a public school serving a majority low-income, Hispanic population. Five children were recruited from a second day camp, hosted at a public school serving a majority low-income, Hispanic and African American population.

The *no ruler* condition included 27 children (M = 8.7 years, SD = .79, 44% female, 96% Hispanic, 4% African American), the *congruent ruler* condition included 27 children (M = 8.7 years, SD = .86, 48% female, 93% Hispanic, 7% African American), and the *incongruent ruler* condition included 26 children (M = 8.5 years, SD = .72, 58% female, 92% Hispanic, 8% African American). Two students who initially began the study, one from the *no ruler* condition and one from the *incongruent ruler* condition, were unable to complete the training. In both cases the children made little progress and voluntarily requested to terminate participation.

3.1.2. Experimental design

Participants were assigned to a condition using a stratified random assignment procedure. Specifically, triads of children from each grade level were randomly assigned to each of three conditions, ensuring that each grade level had a roughly equal number of participants in each condition. To avoid "proactive interference" from a pretest (Opfer & Thompson, 2008), participating children began the main experimental session with the learning task. Children engaged in the learning task until reaching criterion (8 out of 8 correct in a block of trials), and then proceeded immediately to the posttest.

3.1.3 Materials and procedure

3.1.3.1. Standardized measures. To ensure that children from each condition had similar levels of mathematical achievement the Woodcock Johnson III Calculation and Mathematical Fluency subtests were administered in small groups of mixed-condition students in a quiet room. These assessments were chosen because of their previous association with numerical estimation ability (Halberda et al., 2008). Also, these measures provided a potentially interesting covariate with outcome measures.

3.1.3.2. Number line estimation training game. Training was administered one-to-one in either a private room, or in a private area of a large room. The child was placed at a desk with a computer, while the experimenter sat to the child's side to provide assistance. Both testing and training was completed on a Dell laptop with a 17" display. The training software was authored in the Adobe Flash CS4 system and ran as a desktop application using Adobe Air.

At the start of the training game the children viewed a short animated instructional sequence, which provided the narrative context of the game. The children were told that they would be assisting a character "Alex" in his attempt to catch a large number of fish to populate his fish tank. Children viewed an animated image of a map showing a number of different lakes. Each lake was representative of a single block of eight estimation trials. Children were told by the experimenter that they would attempt to catch all eight fish at a lake, and move on to a different lake if they missed one or more fish.

Following the introduction, children were presented with a side view of a lake, with a number line drawn across the surface (30 cm long). The target magnitude was displayed at the top-center of the screen, with the text, "The fish is at [the target] feet". After a second, several of the contextual elements

of the display faded to promote children's attention to the task (see Figure 2). During the first trial the experimenter engaged in condition-specific instruction of the task.

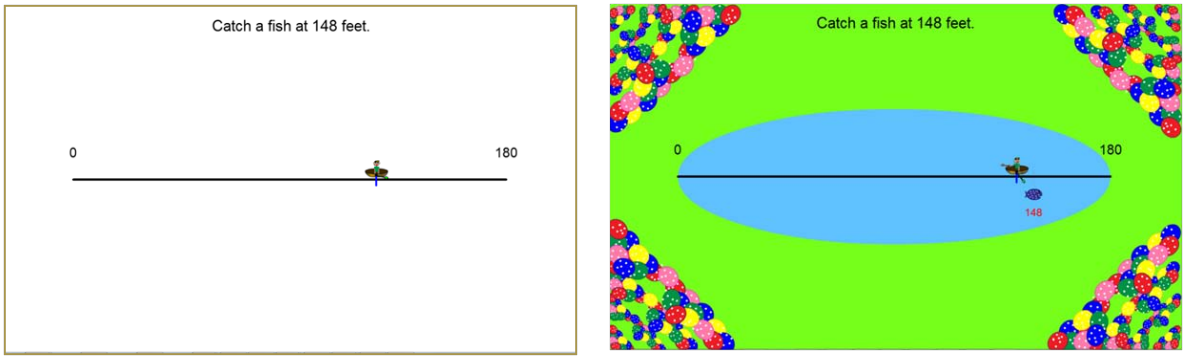


Figure 2. Estimation trial screenshots. Left side shows trial in progress with contextual features removed. Right side shows estimation trial during feedback with contextual features resumed. On the following trial, contextual elements again faded after one second.

In the *no ruler* condition the children were simply told (roughly) that, "The boat starts here at zero. The lake is 180 feet long. The fish is somewhere between 0 and 180, but you have to guess where. The words at the top of the screen give you a hint. You have to think about where that number is between 0 and 180. Can you find the fish?" Following this explanation the children were free to begin the game.

In the *congruent ruler* condition the children received the same explanation of the task as in the *no ruler* condition. However, following this instruction the child was presented with a ruler. In this case the number line printed on the ruler was identical in width (30 cm's) to the number line displayed on screen. On the ruler, accurately placed hatch-marks and their associated values were displayed at 0, 45, 90, 135, and 180. The children were shown how the length of the ruler matched the length of the number line on screen. The child was told that 90 is in the middle, 45 is in the middle between 0 and 90, and 135 is in the middle between 90 and 180. The child was then handed the ruler and asked to demonstrate his or her understanding by pointed to the locations of 90, 45, and 135 on the number line displayed on-screen. If the child did not correctly align the ruler, the experimenter provided corrective feedback. Following this instruction the child was free to begin the game. In some cases, when the child appeared to struggle holding the ruler, the experimenter offered to hold the ruler for the child or place the ruler at the bottom of the screen, in alignment with the target number line.

Finally, in the *incongruent ruler* condition the children received the same explanation as in the other two conditions and were given a ruler, which was described (90 in middle, etc.) as above. However,

in this condition the ruler was constructed to be 33% larger than the *congruent ruler*. The child was told that the ruler was "mistakenly" made too large, and did not match the line on screen. The experimenter placed the ruler on the screen to show this discrepancy. The experimenter then told the child that he or she would need to figure out where the numbers on the ruler could be located on-screen. The child was asked to point to the locations of 90, 45, and 135 on the screen. If the child pointed to visibly inaccurate magnitudes the experimenter provided corrective feedback. Following this instruction the child was free to begin the game. In some cases children were reminded to use the ruler during training trials.

The child then performed a series of eight estimations in a block. Each target magnitude was sampled, at random, from one of eight sub-intervals of the number line (i.e., 1 – 22, 23 – 45, 46 – 68, 68 – 90, 91 – 113, 113 – 135, 136 – 157, and 158 – 180). Each block consisted of one and only one sample from each sub-interval, to ensure that each block had a well-distributed set of targets. During each trial the child navigated the "boat", via horizontal movement of the mouse, and pressed the left mouse button to place his or her final estimate. If the selected magnitude was within 10% of the correct magnitude the child "caught" the fish. He or she was then rewarded with an animated sequence of Alex catching the fish. If the selected magnitude fell out of the 10% margin of error, the actual location of the fish was displayed. The child would then click upon the fish to acknowledge the feedback and proceed to the next trial.

Following a block of eight trials, an animation was displayed to provide a brief respite as well as summary feedback. In this animation the child viewed a number of fish swimming in the fish tank equal to the number that he or she "caught" in the prior block. If the child was able to catch all eight fish training was completed. Otherwise, the child was told that he or she needed to try again to catch all eight fish. After a few moments the child returned to a new level (i.e., with a different background and type of fish) to resume estimation.

The decision to provide a relatively strict training criterion (8 out of 8 in a block) was intended to ensure that the child was familiar with magnitudes distributed across the number line. In pilot testing nearly all children were able to successfully estimate a full block of trials correctly, eventually.

3.1.3.3. Number line estimation posttest. Following successful completion of the training game children immediately began the computerized posttest consisting of four subtests of estimation trials over

four spatially-distinct number lines. Each subtest consisted of 19 trials, whose target magnitudes were sampled from 16 equal sub-intervals of the 0-180 range. Additionally, landmark values of 45, 90, and 135 were included, resulting in the target set: 5, 16, 31, 36, 45, 49, 58, 70, 81, 90, 94, 106, 120, 131, 135, 140, 155, 161, and 178. The software program randomly sorted this set of targets at the start of each subtest.

Each trial displayed a number line with hatch marks and corresponding printed text at 0 and 180 on screen. The target value was printed within a rectangle, with no other text, and displayed at the top-center of the screen. Upon initiating a trial a small green triangle was displayed over the zero. Participants were asked to click on this green triangle to begin the trial. Upon clicking, the triangle was replaced with a moveable blue hatch mark that could be shifted across the length of the number line. When the participant was ready to provide the final estimate, he or she simply pressed down on the mouse until the blue hatch mark turned red. A "pressing" action was required (i.e., greater than 500 msec.), rather than a click, to prevent unintended final estimates.

If the child was satisfied with his or her final estimate the experimenter pressed the space key to continue to the next trial. If the child immediately recognized that he or she had made a mistake the experimenter could press the delete key to place the trial back in the randomized queue of subtest trials. In some cases, if the child appeared to accidentally press the mouse button or seemed to be inattentive to the task the experimenter would ask, "Is that where you wanted to put it?" If the child said "no" the experimenter would press delete, otherwise the experimenter would continue to the next trial. No feedback was provided by the computer. The experimenter provided only general support, such as "You're doing great, keep it up."

On the first subtest the length of the number line was identical to the training task (30 cm). In this case, the child was also told that the task was similar to the fishing game. The second number line, also 0-180, was half the length (15 cm, see Table 1). The third number line was the same length as the first number line (30 cm), but displayed "in reverse" (i.e., 0 on the right, 180 on the left). The fourth number line was the same length as the second number line (15 cm), but oriented vertically (i.e., 0 on the bottom, 180 on the top). In each of the last three subtests the child was asked to explain, prior to estimation, how the number line was different. If the child did not notice the critical difference, this feature was explained.

Table 1
Spatial design of number lines in training and testing

	Training	Posttest			
		Subtest 1	Subtest 2	Subtest 3	Subtest 4
Length	30 cm	30 cm	15 cm	30 cm	15 cm
Orientation	Horizontal, left-to-right	Horizontal, left-to-right	Horizontal, left-to-right	Horizontal, right-to-left	Vertical, bottom-to-top

Finally, upon completing all four subtests the child was asked to verbally state the midpoint, 1st quartile, and 3rd quartile of the 0-180 range, respectively. Specifically, the experimenter stated, "On all of those number lines zero was on one side and one hundred-eighty was on the other. What number would go right in the middle?" After answering the question, correctly or incorrectly, the experiment stated, "Imagine that we had a number line that goes from zero on one side to ninety on the other. What number would go in the middle?" Finally, the latter question would be repeated in the context of a number line ranging from 90 to 180. Children were not given a visual representation of a number line for this task. Answers were recorded on a computer spreadsheet.

3.2. Results

3.2.1. Standardized measures. Participants in the *no ruler* condition received a mean standardized score, grade-normed, of 95.7 (SD = 12.2) on Math Fluency and 101.4 (SD = 10.3) on Calculation. Participants in the *congruent ruler* condition received a mean standardized score, grade-normed, of 97.0 on Math Fluency (SD = 11.7) and 104.7 on Calculation (SD = 9.6). Participants in the *incongruent ruler* condition received a mean standardized score, grade-normed, of 102.2 (SD = 9.6) on Math Fluency and 107.5 on Calculation (SD = 9.2). An ANOVA with condition as a between-participants factor did not indicate a significant difference between conditions [Calculation: $F[2, 77] = 2.4$, $p = .1$; Math Fluency: $F[2, 77] = 1.0$, $p > .1$].

3.2.2. **Number line estimation training game**. All 80 children were able to complete the learning task by reaching criterion. However, the duration of the task and number of trials (or blocks)

differed between conditions. As a summary, in Figure 3, both total number of blocks to reach criterion and total duration (not including animated introduction or feedback) were plotted for each of the three conditions.

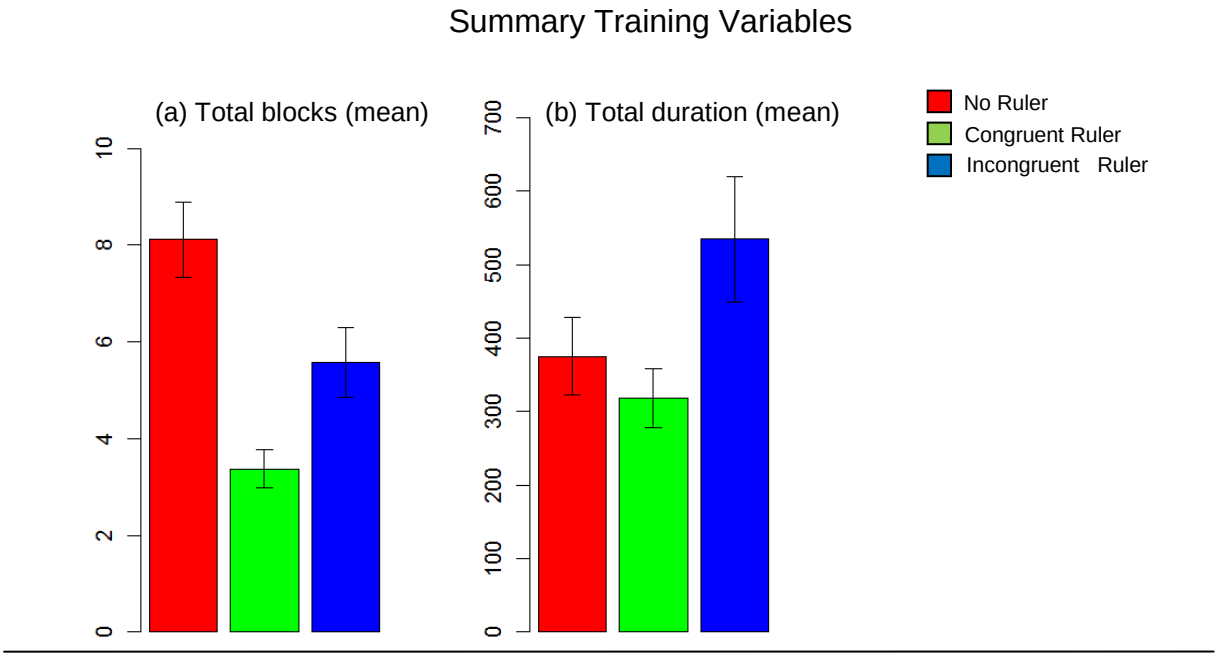


Figure 3. Summary training variables: (a) The number of blocks needed to reach criterion. (b) The total duration for the entire training session (in seconds).

As expected, an ANOVA of "total blocks" revealed a significant effect of condition [$F(2, 77) = 13.5$, $p < .001$, $\eta_p^2 = .26$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 19.1$, $p < .001$, $\eta_p^2 = .25$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 7.5$, $p = .008$, $\eta_p^2 = .10$].

In addition to the effect of condition, Figure 4 displays the relationship between total blocks and age. Overall, total blocks and age were not significantly correlated [$r(78) = -.10$, $p = .40$]. Likewise, for children in the *no* ruler condition total blocks and age were not significantly correlated [$r(25) = .09$, $p = .64$]. On the other hand, for children in the *congruent ruler* condition total blocks and age were significantly correlated [$r(25) = -.43$, $p = .03$]. For children in the in*congruent ruler* condition total blocks and age showed a trend towards a significant correlation [$r(24) = -.33$, $p = .10$].

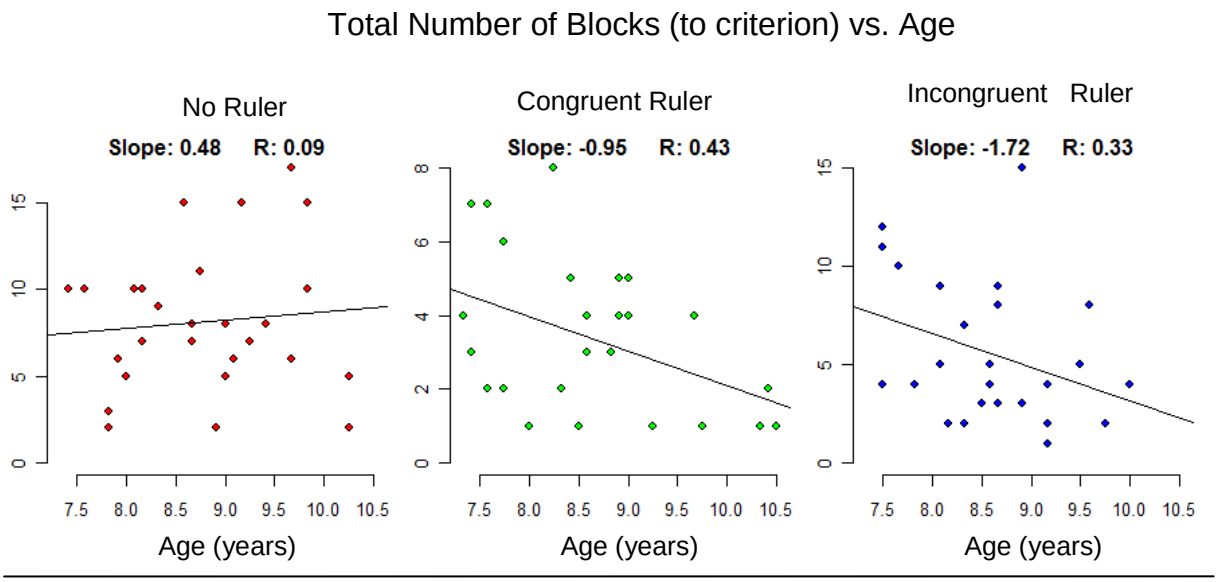


Figure 4. Each graph displays all participants in a condition. The x-axis represents age (in years). The y-axis represents the number of blocks needed to reach criterion for the corresponding participant.

Utilizing the limited moderating effects of age on total blocks, an ANCOVA of total blocks revealed a significant effect of condition [$F(2, 77) = 14.9$, $p < .001$, $\eta_p^2 = .28$] and a significant effect of age [$F(1, 77) = 5.3$, $p = .02$, $\eta_p^2 = .07$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 20.0$, $p < .001$, $\eta_p^2 = .21$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 9.5$, $p = .003$, $\eta_p^2 = .11$]. Appendix A displays ANOVA and ANCOVA tables related to total blocks.

Beyond differences in the mean number of blocks needed to reach criterion, Figure 5 reveals somewhat differing patterns of participants reaching criterion at each block for each condition. Specifically, for the *congruent ruler* condition at least one child reached criterion in the first 8 blocks of training. Furthermore, by the fifth block 20 of 27 participants had reached criterion. On the other hand, in the *no ruler* condition, by the fifth block only 4 of 27 participants had reached criterion. While the number of participants who reached criterion, in the *no ruler* condition, continued to increase rapidly following the fifth block, by the twelfth block – well beyond the number of blocks needed by any subject in the *congruent ruler* condition – four participants still remained, and needed at least three more blocks to complete training. Children in the *incongruent ruler* condition appeared to follow a middle path.

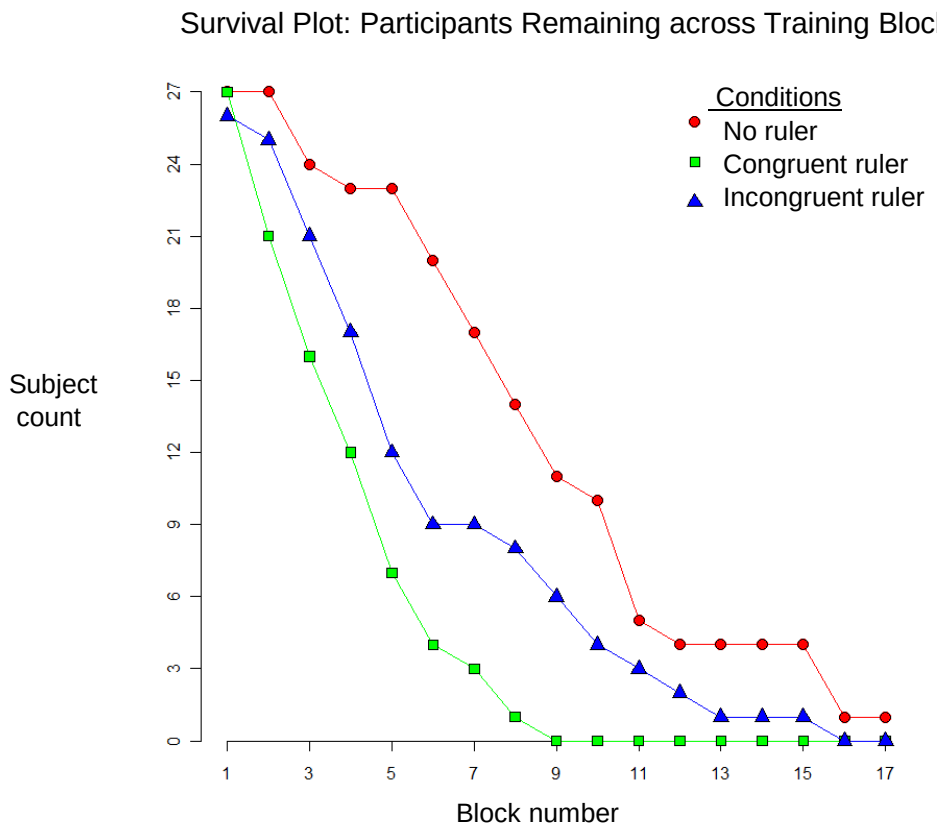


Figure 5. Tracks the number of participants currently participating across all training blocks. For the *no ruler* condition the final subject completed training on the 17th block. For the *congruent ruler* the final subject completed training on the 8th block. For the *incongruent ruler* the final subject completed training on the 15th block.

For total duration (Figure 3b) the pattern of means was dissimilar to that of total blocks (Figure 3a). Specifically, children in the *incongruent ruler* condition appeared to have taken longer to reach criterion than children in the *no ruler* condition, in spite of having completed fewer blocks. An ANOVA of total duration revealed a significant effect of condition [$F(2, 77) = 3.3$, $p = .04$, $\eta_p^2 = .08$]. Planned comparisons showed a trend towards a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 3.3$, $p = .07$, $\eta_p^2 = .04$], as well as a trend towards a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 3.3$, $p = .07$, $\eta_p^2 = .04$]. Post-hoc comparisons, with Bonferroni-adjusted alpha levels, showed only a significant difference between participants in the *congruent ruler* condition and the *incongruent ruler* condition [$p = .05$].

Figure 6 displays the relationship between total duration and age, by condition. Overall, total duration and age were significantly correlated [$r(78) = -.37$, $p < .001$]. For children in the *no* ruler

condition total duration and age showed a trend towards significant correlation [r(25) = -.34, p = .08]. Similarly, for children in the *congruent ruler* condition total blocks and age showed a trend towards significant correlation [r(25) = -.34, p = .08]. Finally, for children in the in*congruent ruler* condition total blocks and age were significantly correlated [r(24) = -.49, p = .01].

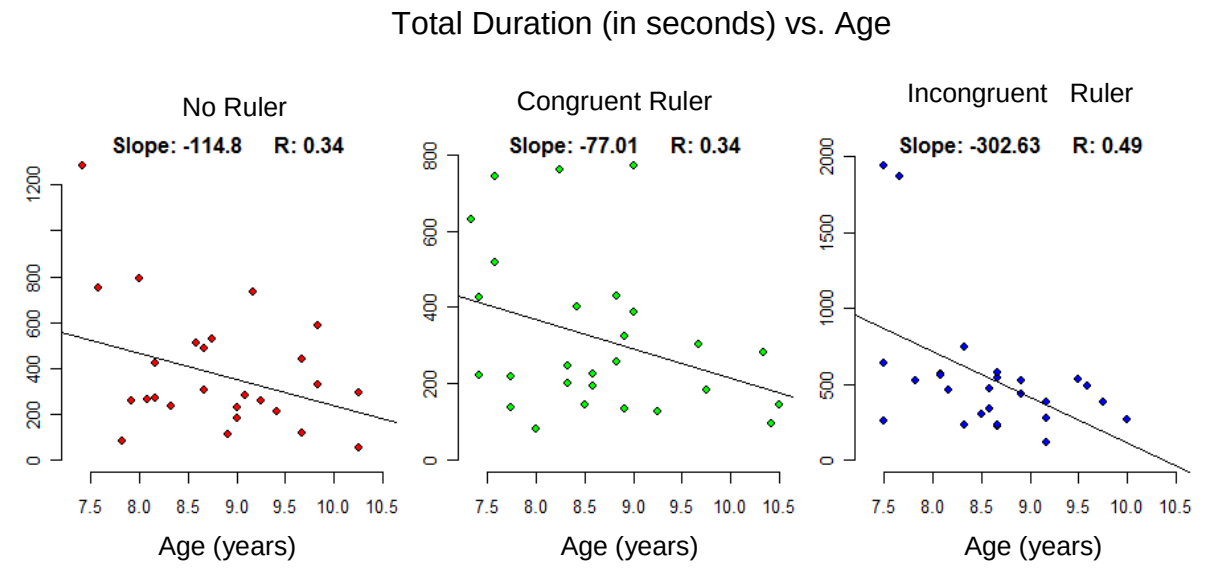


Figure 6. Each graph displays all participants in a condition. The x-axis represents age (in years). The y-axis represents the number of seconds spent in estimation trials for the corresponding participant.

Utilizing the moderating effects of age, an ANCOVA of total duration revealed a trend towards a significant effect of condition [$F(2, 77) = 3.0$, $p = .06$, $\eta_p^2 = .07$] and a significant effect of age [$F(1, 77) = 13.9$, $p < .001$, $\eta_p^2 = .15$]. Due to the lack of a strong relationship between condition and total duration, when accounting for age of participants, no further planned comparisons or post-hoc tests were conducted.

The divergent findings between total blocks and total duration – which would be expected to be highly related – likely reflects distinct contributions of treatment to both the initial instructional time and the application of estimation strategies. Figure 7 displays the mean duration of the first 4 training blocks, separated by condition, as well as the average of each subject's mean duration across his or her completed blocks. For example, if student 1 completed three blocks in 100 seconds, 200 seconds, and 300 seconds, his mean duration would be 200 seconds. If a second student completed two blocks in 200 seconds and 100 seconds, her mean duration would be 150 seconds. The mean of these means would

then be 175 seconds. While this "mean of means" is clearly confounded by the number of blocks that the child has performed (i.e. with less blocks, longer initial blocks will produce elevated means), it is a determining factor of total duration, along with number of blocks, and therefore worth display.

As the graphs reveals, the average duration (and therefore total duration) were highly influenced by the first block, where the duration, for each condition, was more than double the duration of the second block. For all three conditions, among children that remained until (at least) block 2, participants' durations from block 1 to block 2 dropped significantly [*no ruler:* $t(26) = 8.2$, $p < .001$; *congruent ruler:* $t(20) = 7.0$, $p < .001$; *incongruent ruler:* $t(25) = 10.2$, $p < .001$].

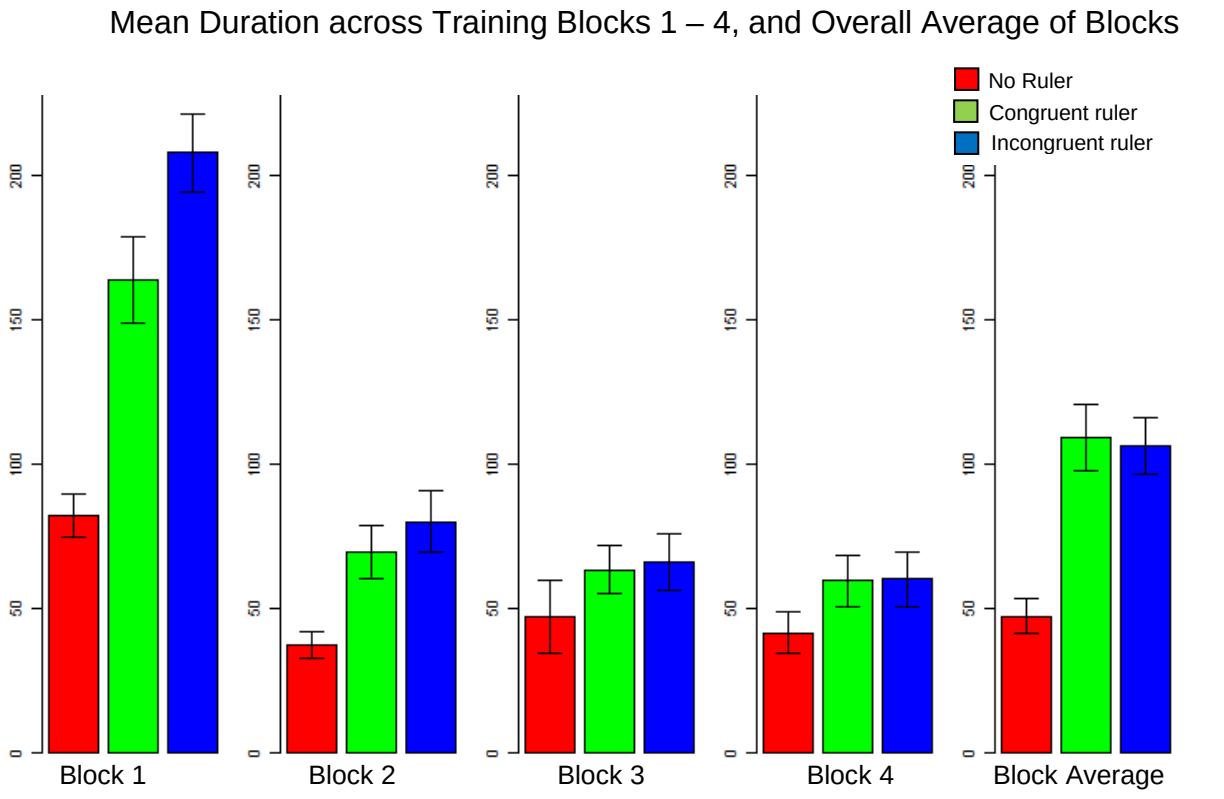


Figure 7. The mean duration (in seconds) of subject in the first four blocks in training. Block 1 clearly reflects the influence of initial instruction. The final graph shows the overall mean of each child's average block duration. Although the congruent ruler average duration was slightly (non-significantly) higher than the incongruent ruler average, this likely reflected a greater influence of block 1 for congruent ruler participants.

As an indication of the effects of differing instructional times (primarily), an ANOVA of block 1 mean duration revealed a significant effect of condition [$F(2, 77) = 26.4$, $p < .001$, $\eta_p^2 = .41$]. Post-hoc comparisons, with Bonferroni-adjusted alpha levels, revealed a significant difference between children in

the *no ruler* condition and the *congruent ruler* condition [$p < .001$], a significant difference between children in the *no ruler* condition and the in*congruent ruler* condition [$p < .001$], and a significant difference between children in the *congruent ruler* condition and the *incongruent ruler* condition [$p < .05$]. Thus, in this case there is a clear hierarchy in block 1 durations such that children in the *no ruler* condition, on average, completed the block fastest, while children in the *incongruent ruler* condition completed the block, on average, slowest.

On following blocks, which did not contain instruction, differences between conditions likely reflected effects of diverging strategies. However, six children in the *congruent ruler* condition and one child in the *incongruent ruler* condition completed training in the first block. Therefore, differences between conditions may reflect a loss of these high-performing children. With this in mind, an ANOVA of block 2 mean duration revealed a significant effect of condition [$F(2, 77) = 7.4$, $p = .001$, $\eta^2 = .17$]. Post-hoc comparisons, with Bonferroni-adjusted alpha levels, revealed a significant difference between children in the *no ruler* condition and the *congruent ruler* condition [$p = .01$], a significant difference between children in the *no ruler* condition and the in*congruent ruler* condition [$p < .001$], but no difference between children in the *congruent ruler* condition and the *incongruent ruler* condition [$p > .1$]. Given the strength of the difference between the *no ruler* and the *incongruent ruler* condition, the loss of one subject is not likely to have made a strong impact. The effect of losing 6 participants from the *congruent ruler* condition is unclear.

While the effects of total duration and total blocks to reach criterion reflect, indirectly, the role of estimation accuracy, a more direct measure of accuracy is appropriate. Specifically, the mean percent absolute error (or mean PAE), measures the absolute difference between the estimated and actual magnitude divided by the range of the estimation scale (180 here). For example, presented with a target of 90, an estimate of 81 would produce a PAE of 5% [(90 – 81) / 180 * 100%]. Figure 8, displays the differences between conditions in mean PAE for the first 4 blocks, and the overall average.

In the case of the *no ruler* condition, where the number of participants was constant from block 1 to block 2, participants' significantly reduced errors from block 1 to block 2 [$t(26) = 2.7$, $p = .01$]. In the case of the *congruent ruler* condition, the 21 (of 27) participants that remained until (at least) block 2 showed no significant reduction in errors [$t[20] = 1.3$, $p = .21$]. Finally, in the case of the *incongruent ruler*

condition, the 25 (of 26) participants that remained showed a trend towards a significant reduction of errors [t(24) = 2.0, p = .06].

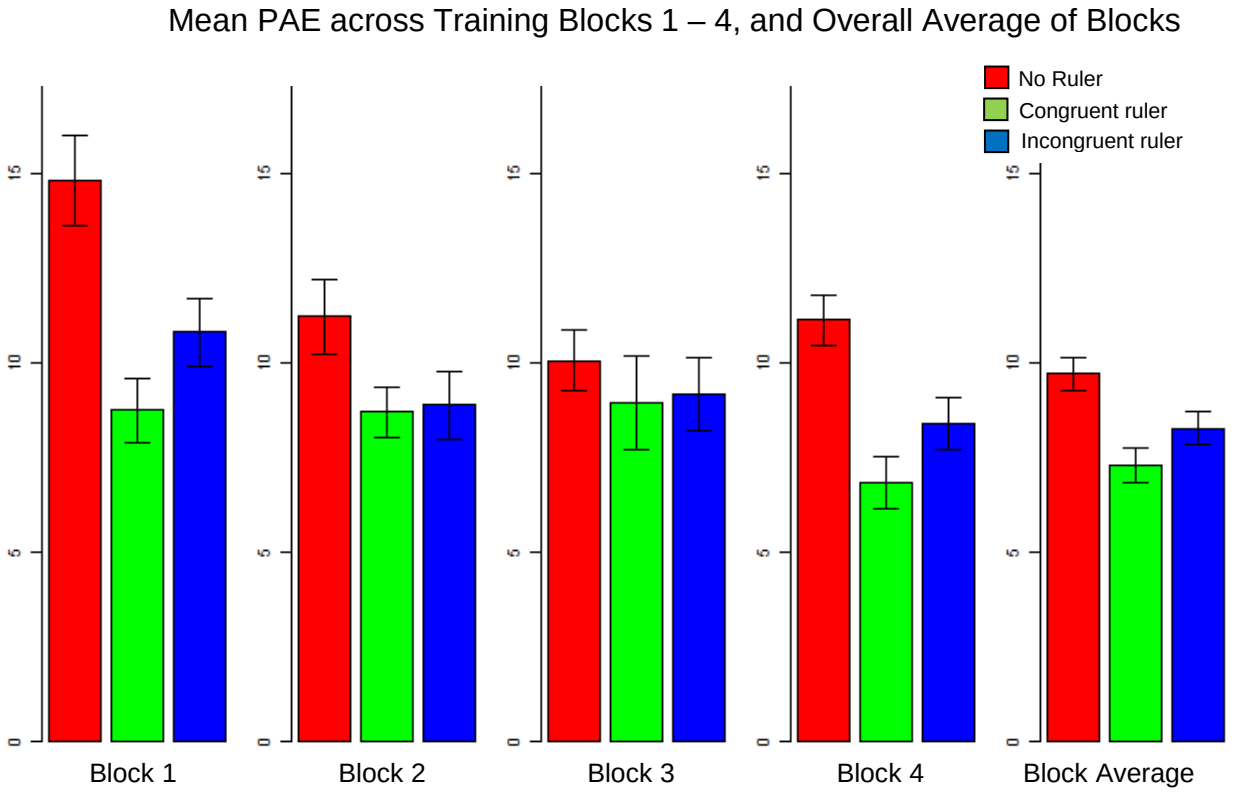


Figure 8. The mean PAE of participants in the first four blocks in training. The final graph shows the overall average mean PAE.

To test between-subject differences in the first block, an ANOVA of mean PAE revealed a significant effect of condition [$F(2, 77) = 9.8$, $p < .001$, $\eta_p^2 = .20$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 10.5$, $p = .001$, $\eta_p^2 = .13$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 6.8$, $p < .005$, $\eta_p^2 = .10$]. On the other hand, an ANOVA of mean PAE at block 2 (with 71 remaining participants) revealed only a trend towards a significant effect of condition [$F(2, 70) = 2.7$, $p = .08$, $\eta_p^2 = .07$]. Further contrasts and post-hoc tests were not significant.

The final column of bars in Figure 8 displays an average, across all blocks, of individuals mean PAE's. For example, if student 1 completed three blocks with an average 10% mean PAE, 20% mean PAE, and 30% seconds, his overall mean PAE would be 20%. If a second student completed two blocks

with 20% mean PAE and 10% mean PAE, her overall mean PAE would be 15%. The mean of these means would then be 17.5% error. Because this statistic was gathered from a differing number of blocks, by condition, an ANOVA would be affected by the diverging average number of blocks between conditions. Nonetheless, an ANOVA table is displayed in Appendix A, demonstrating similar relationships to the analysis of block 1.

3.2.3. Relationships between individual training performance and posttest performance

The summary measures, described in the previous section, vary by condition, also by age (in some cases). The effect of age (or possibly as a proxy for grade), likely reflected the prior numerical experience of the participants. Older children are more likely to have engaged in numerical tasks with numbers up to and exceeding 100. Beyond effects of age, individual differences in training performance likely reflected differences in numerical ability. As such, we should expect some relationship between training performance and testing performance.

The relationship between training and testing is displayed for each subject in Appendix B. Specifically, each plot displays the mean PAE of every testing block and post-subtest for an individual subject. The difference between conditions is most evident in the number of testing points (*no ruler* with the most and *congruent ruler* with the least). While it is difficult to discern definitive patterns in these graphs, there appears to be an inverse relationship between the number of and error (height) of training points and the error of testing points.

These relationships between posttest mean PAE and training measures are summarized in Table 2, below. The mean PAE of the first training block (8 trials) correlated highly with performance in the *congruent ruler* condition, and not with other conditions. The overall training mean PAE for participants (i.e., the average of block means) showed a significant correlation with all conditions in at least 2 of four subtests. The total blocks measure showed a significant correlation with the *no ruler* condition only in the first subtest, with the *congruent ruler* condition in the second and third subtests, and with the *incongruent ruler* condition in the second subtest, only. Finally, total duration only showed a significant correlation with the *no ruler* condition in the second subtest, the *congruent ruler* condition in the second and third subtests, and not at all with the *incongruent ruler* condition. These analyses demonstrate that greater time-on-task, within a condition, was associated with poorer performance at posttest.

Table 2

Training summary measures vs. testing mean PAE

Subtest #	Condition	Training Measures			
		Block 1 mean PAE	Overall mean PAE	Total blocks	Total duration
1	No ruler	.37 †	.70 ***	0.42*	0.36†
	Congruent	.44 *	.37 †	0.27	0.33†
	Incongruent	.28	.58 **	0.16	0.27
2	No ruler	.21	.51 **	0.26	0.53**
	Congruent	.56 **	.40 *	0.42*	0.47*
	Incongruent	.22	.41 *	0.42*	0.34†
3	No ruler	.23	.41 *	0.08	0.18
	Congruent	.70 ***	.54 **	0.42*	0.32
	Incongruent	.01	.33	0.35†	0.08
4	No ruler	.14	.36 †	0.08	0.15
	Congruent	.48 *	.45 *	0.31	0.39*
	Incongruent	.14	.32	0.36†	0.08

*** p < .001 ** p < .01 * p < .05 † p < .1

3.2.4. *Posttest accuracy*

In Figure 9, below, an example of one child's posttest estimates for a single subtest is displayed. In her case the data demonstrated a logarithmic relationship between estimated and actual magnitude (best fit by the red logarithmic line). In previous studies of numerical estimation (e.g. Siegler & Opfer, 2003), median estimates are used to demonstrate group differences in overall model (i.e., logarithmic or linear). Surprisingly there were no such qualitative differences between groups at posttest. In all groups, across all four tests, a linear model fit median estimates better than a logarithmic model (Appendix C-1). Yet, this does not entail that all children's individual estimates were best fit by a linear model. Individual estimate graphs (Appendix D) provide evidence for logarithmic representations, particularly with younger children. To show this, children were divided into two groups by median age to produce separate plots for older and younger children (Appendices C-2, C-3). While only one of these graphs shows clear

evidence of a logarithmic model (subtest 3, *congruent ruler* condition), overall lower slopes and variance explained (R^2) suggest that many of these younger children tended towards logarithmic performance.

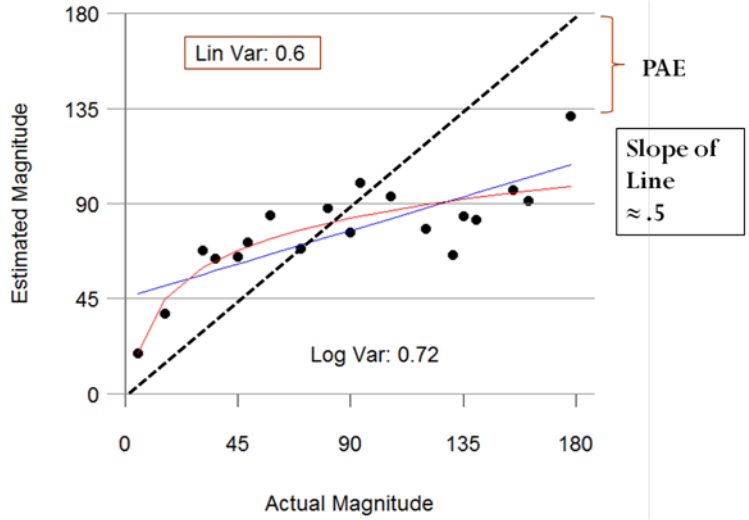


Figure 9. Example of a logarithmic distribution of estimates – i.e. the variance accounted for by the linear model (.6) was less than the variance accounted for by the logarithmic model (.72). PAE is defined as the distance from the dashed, diagonal line. Slope is defined as the slope of the linear (blue) line.

In previous studies of number line estimation interventions (Ramani & Siegler, 2008; Siegler & Ramani, 2008; Siegler & Ramani, 2009) three primary measures of accuracy – i.e., linearity, slope, and mean percent absolute error – were used to assess the quality of children's estimates (see Figure 9 also provides a visual depiction of each). Linearity refers to the amount of variance explained by the best-fitting linear function to estimated magnitudes. Slope refers to the slope of this best-fitting linear regression line. Finally, the mean percent absolute error (PAE) refers to the mean percent difference between estimated and actual magnitudes. For example, in the training trials, the PAEs of successful trials are, by definition, less than 10%.

Table 3, shown below, displays the relationship between these three outcome measures and measures of individual differences, including standardized scores, age, and grade level – revealing generally significant correlations between variables. Individual differences appeared to influence mean PAE the most. Of the sixteen correlations with mean PAE, only one did not show a significant correlation (with $\alpha < .05$). Of the four measures of individual difference only age showed a significant correlation with all outcome measures, at each subtest – thereby making it an ideal candidate as a covariate.

Table 3

Testing outcome vs. individual difference measures correlation table

Posttest Measure	Subtest #	Age	Grade	WJ Calc SS	WJ Fluency SS
Mean PAE	1	-.31 **	-.26 **	-.39 ***	-.26 *
	2	-.38 ***	-.41 ***	-.32 **	-.32 **
	3	-.27 *	-.30 **	-.22 *	-.18
	4	-.36 ***	-.24 *	-.33 **	-.27 *
Linearity	1	.23 *	.14	.42 ***	.33 **
	2	.23 *	.30 **	.27 *	.32 **
	3	.28 *	.29 **	.15	.17
	4	.23 *	.09	.21 †	.25 *
Slope	1	.23 *	.17	.31 **	.21 †
	2	.34 **	.38 ***	.19 †	.26 *
	3	.27 *	.29 **	.15	.13
	4	.28 *	.19 †	.24 *	.22 †

*** p < .001 ** p < .01 * p < .05 † p < .1

Although the relationships between variables, shown in Table 3, were consistently significant, the strengths of the correlations were only moderate. As such, correlations were less likely to be significant when separated by treatment groups. However, since age was used as a covariate in analyses of outcome measures, the relationship between age and subtest outcome score, separated by condition, are plotted in Appendix E.

Turning towards specific hypotheses regarding posttest accuracy, Figure 10, shown below, displays a general trend towards increased error as a function of subtest, as well as a relatively static relationship between conditions, such that errors in the *congruent ruler* condition were greater than error in the *no ruler* condition. Errors in the *no ruler* condition were, in turn, greater than the errors in the *incongruent ruler* condition – as predicted. An ANOVA of mean PAE revealed a significant effect of condition [$F(2, 77) = 8.3$, $p = .001$, $\eta_p^2 = .18$], a significant effect of subtest [$F(3, 231) = 15.9$, $p < .001$, $\eta_p^2 = .17$], but no significant interaction between condition and subtest [$F(6, 231) = 1.7$, $p = .12$, $\eta_p^2 = .04$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 12.5$, $p < .001$, $\eta_p^2 = .13$], as well as a

significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [F(1, 77) = 5.1, p < .05, η_p^2 = .05].

Utilizing the moderating effects of age, an ANCOVA of mean PAE revealed a significant effect of condition [F(2, 76) = 12.2, p < .001, η_p^2 = .24], a significant effect of age [F(1, 76) = 21.6, p < .001, η_p^2 = .22], a significant effect of subtest [F(3, 228) = 16.1, p < .001, η_p^2 = .18], a trend towards a significant interaction between condition and subtest [F(6, 228) = 1.9, p = .09, η_p^2 = .05], and no significant interaction of age and subtest [F(3, 228) = 1.9, p = .13, η_p^2 = .03]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [F(1, 76) = 16.2, p < .001, η_p^2 = .17], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [F(1, 76) = 9.5, p = .003, η_p^2 = .11].

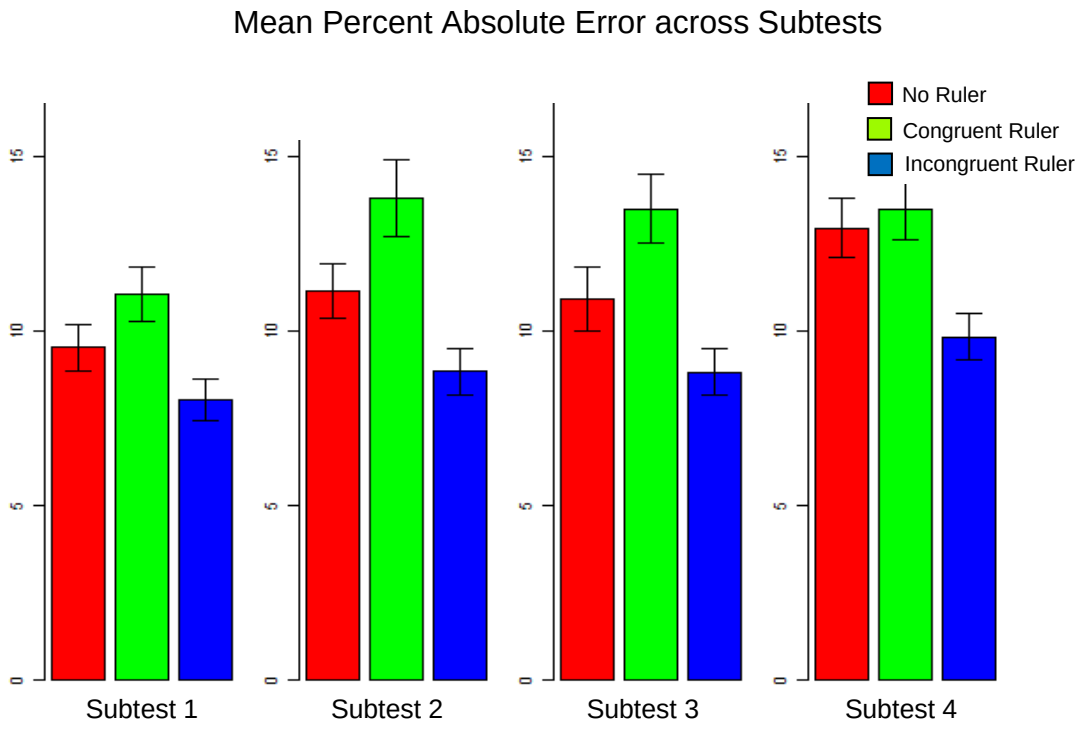


Figure 10. Mean percent absolute error for all participants, across four subtests.

Likewise, participants demonstrated a general trend towards decreased linearity (R^2 of the linear model fitting estimated magnitude to actual magnitude) as a function of subtest (Figure 11). The general relationships between conditions were similar to those described for mean PAE. An ANOVA of linearity

revealed a significant effect of subtest [$F(3, 231) = 9.4$, $p < .001$, $\eta_p^2 = .11$], a significant effect of condition [$F(2, 77) = 7.3$, $p = .001$, $\eta_p^2 = .16$], but no significant interaction between condition and subtest [$F(6, 231) = 1.3$, $p = .27$, $\eta_p^2 = .03$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 10.0$, $p < .01$, $\eta_p^2 = .13$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 4.7$, $p < .05$, $\eta_p^2 = .06$].

Utilizing the moderating effects of age, an ANCOVA of linearity revealed a significant effect of condition [$F(2, 76) = 9.4$, $p < .001$, $\eta_p^2 = .20$], a significant effect of age [$F(1, 76) = 11.8$, $p = .001$, $\eta_p^2 = .14$], a significant effect of subtest [$F(3, 228) = 9.3$, $p < .001$, $\eta_p^2 = .11$], no significant interaction between condition and subtest [$F(6, 228) = 1.3$, $p = .25$, $\eta_p^2 = .03$], and no significant interaction of age and subtest [$F(3, 228) = .42$, $p = .74$, $\eta_p^2 = .01$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 76) = 11.6$, $p = .001$, $\eta_p^2 = .13$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 76) = 6.4$, $p = .008$, $\eta_p^2 = .09$].

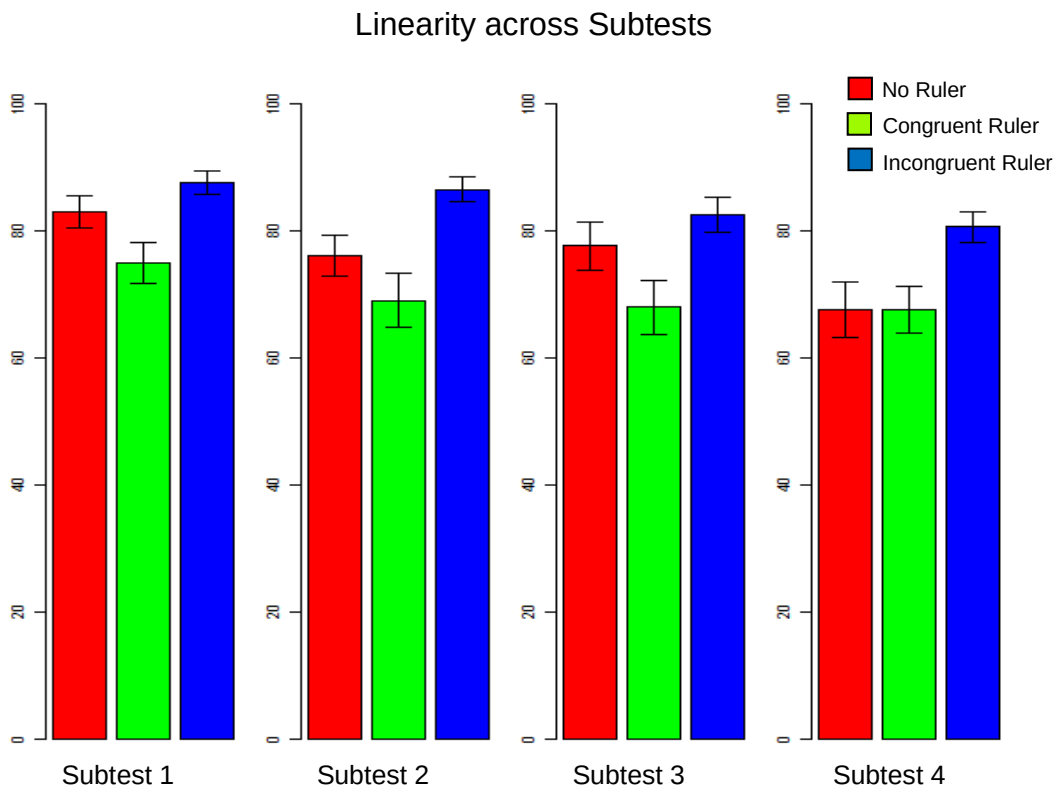


Figure 11. Mean linearity for all participants across all four tests.

Finally, participants demonstrated a general trend towards decreased slope (of the linear fitting model) as a function of subtest (Figure 12). The general relationships between conditions were similar to those described in mean PAE and linearity, above. An ANOVA of slope revealed a significant effect of subtest [$F(3, 231) = 29.0$, $p < .001$, $\eta_p^2 = .27$], a significant effect of condition [$F(2, 77) = 6.3$, $p < .01$, $\eta_p^2 = .14$], but no significant interaction between condition and subtest [$F(6, 231) = 1.5$, $p = .18$, $\eta_p^2 = .04$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 77) = 10.0$, $p < .01$, $\eta_p^2 = .13$]; however, in this case there was no significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 77) = 2.7$, $p = .11$, $\eta_p^2 = .03$].

Utilizing the moderating effects of age, an ANCOVA of slope revealed a significant effect of condition [$F(2, 76) = 8.3$, $p = .001$, $\eta_p^2 = .18$], a significant effect of age [$F(1, 76) = 13.8$, $p < .001$, $\eta_p^2 = .15$], a significant effect of subtest [$F(3, 228) = 29.1$, $p < .001$, $\eta_p^2 = .28$], no significant interaction between condition and subtest [$F(6, 228) = 1.6$, $p = .16$, $\eta_p^2 = .04$], and no significant interaction of age and subtest [$F(3, 228) = 1.4$, $p = .24$, $\eta_p^2 = .02$]. Planned comparisons showed a significant difference between

children in the *congruent ruler* condition and those children in either of the other two conditions [F(1, 76) = 11.8, p = .001, η_p^2 = .13], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [F(1, 76) = 5.0, p = .03, η_p^2 = .06].

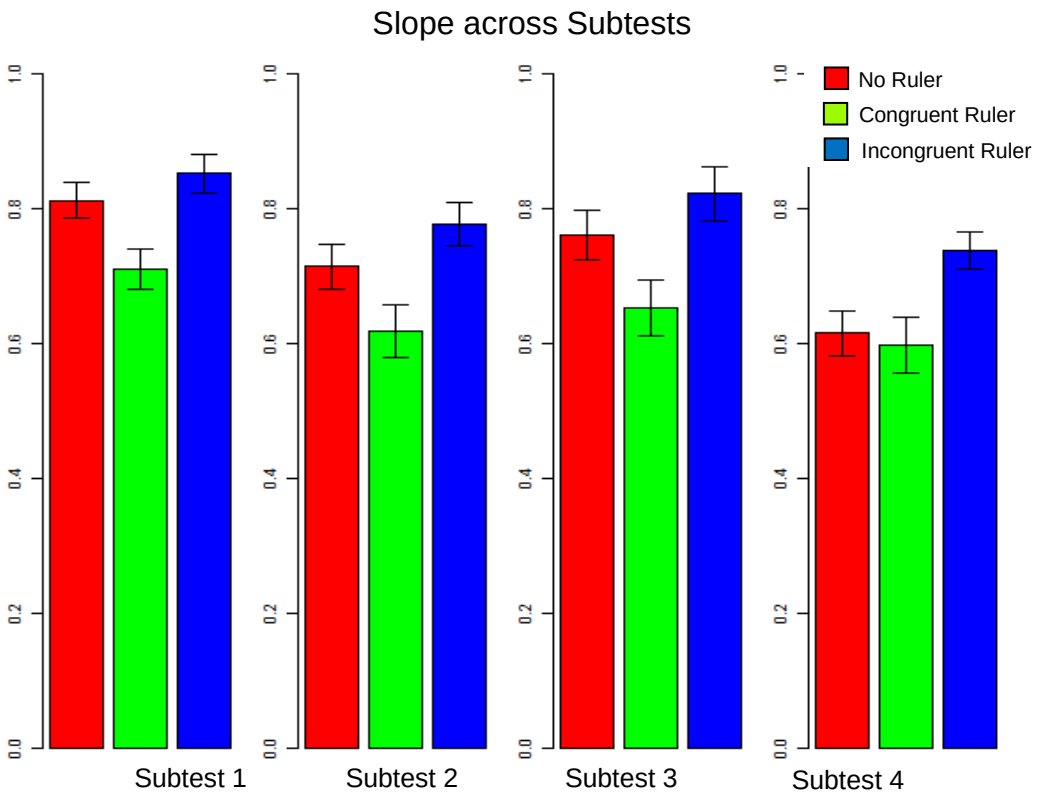


Figure 12. Mean slope for all participants across all four subtests.

The lack of a significant difference between the slopes of estimates in the *no ruler* and the *incongruent ruler* conditions, without the effects of a moderator removed, is perhaps best explained by the estimated over actual magnitudes plots (Appendices C and D). Inspection of the median estimate graphs of participants in the *no ruler* condition reveals that, in general, the higher magnitude estimates appear to level off, resulting in decreased linearity, higher mean PAE, and lower slope. However, in the case of the largest estimate – 178 – this trend is bucked – such that median estimates were highly accurate. While, the contribution of a single data point is unlikely to significantly affect error or variance results – the position of this data point may have enacted significant leverage upon slope. While this may be the case

across all conditions, the effect upon the *no ruler* appears particularly pronounced. To a lesser extent this effect may be found in the judgment of the smallest target, 5, as well.

Anecdotally, upon receiving the target magnitude of 5 or 178, many of the children would state, "that's right near 0" (or 180), whereas on other trials no such exclamation would be made (except for landmark values in ruler conditions). As statistical evidence of 178's unique status in the *no ruler* condition, across all four subtests, Cook's d values are greatest at this magnitude (subtest 1: d = .34; subtest 2: d = .80; subtest 3: d = .47; subtest 4: d = .81). In the case of the 2nd and 4th subtests (where the number line was scaled by 50%), Cook's d values were more than twice as large as the next largest value (subtest 2: d = .37; subtest 4: d = .37). For the *congruent ruler* condition the estimates of 5 had the highest or second highest value of Cook's d. Appendix G displays plots of Cook's d.

Given the potential outlier status of estimates of 5 and 178 an ANOVA on the slope of estimates, with targets 5 and 178 removed, revealed a significant effect of condition [$F(2, 77) = 6.8$, $p = .002$, $\eta_p^2 = .15$] and subtest [$F(3, 231) = 24.6$, $p < .001$, $\eta_p^2 = .24$], and no significant interaction of condition and subtest [$F(6, 77) = 1.38$, $p = .22$, $\eta_p^2 = .04$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions $F(1, 77) = 9.3$, $p = .003$, $\eta_p^2 = .11$], as well as a significant difference between children in the *no ruler* condition and children in the *incongruent ruler* condition [$F(1, 77) = 4.3$, $p = .04$, $\eta_p^2 = .05$], where this was not the case with trials 5 and 178 included.

Utilizing the moderating effects of age, an ANCOVA of slope, with targets 5 and 178 removed, revealed a significant effect of condition [$F(2, 76) = 8.3$, $p = .001$, $\eta_p^2 = .18$], a significant effect of age [$F(1, 76) = 8.6$, $p = .004$, $\eta_p^2 = .10$], a significant effect of subtest [$F(3, 228) = 24.6$, $p < .001$, $\eta^2 = .25$], no significant interaction between condition and subtest [$F(6, 228) = 1.4$, $p = .20$, $\eta_p^2 = .04$], and no significant interaction of age and subtest [$F(3, 228) = 1.0$, $p = .39$, $\eta_p^2 = .01$]. Planned comparisons showed a significant difference between children in the *congruent ruler* condition and those children in either of the other two conditions [$F(1, 76) = 10.4$, $p = .002$, $\eta_p^2 = .12$], as well as a significant difference between children in the *no ruler* condition and the *incongruent ruler* condition [$F(1, 76) = 6.7$, $p = .013$, $\eta_p^2 = .08$]. ANOVA tables for slope, with 5 and 178 removed, are presented in Appendix F. Furthermore, a similar analysis is presented with only 178 removed, showing nearly equivalent results.

Across all of the results discussed here a surprising result, based on initial hypotheses, was the lack of an interaction between subtest and condition (i.e., the only sizable effect was a trend in the mean PAE results). Specifically, while a general decrease in performance across the four subtests was expected, the spatially-generic landmark information, supported by the *incongruent ruler* condition, was intended to buffer against this decay. In support of this notion, the graphs displayed in Figures 10 – 12 suggest that there is greater difference between conditions in subtest 4 then subtest 1. To test this directly, post-hoc t-tests, with Bonferroni-adjusted alpha levels, were conducted to compare mean PAEs between conditions in both subtest 1 and subtest 4 (3 comparisons for each subtest). At subtest 1 the only significant difference was between participants in the *congruent ruler* condition and the *incongruent ruler* condition [$t(51) = 3.1$, $p = .01$]. On the other hand, for block 4 there was a significant difference between the *incongruent ruler* condition and both the *congruent ruler* condition [$t(51) = 3.3$, $p = .005$] and the *no ruler* condition [$t(51) = 2.9$, $p = .02$].

3.2.5. Individual model comparisons

As described above a young child, with an immature concept of numerical magnitude is likely to produce a logarithmic pattern of estimated magnitudes. In contrast, older or relatively high-performing child is more likely to produce a linear pattern of estimates. To assess these patterns blocks of data may be categorized by fitting a logarithmic or linear model to the estimates and selecting the model which accounts for greater variance (Siegler & Opfer, 2003). The distribution of models within a group (i.e., experimental condition) may then provide a measure of overall ability for a group. Group distributions may then be compared directly using Chi-squared tests of independence.

Applying this analysis, each participant's subtest data was fit with both a linear and logarithmic model. Appendix D displays the subtest data for each participant along with the best-fitting model – which included linear, logarithmic, exponential, segmented linear models, and cyclical power models (discussed below). Table 6 displays the distribution of linear and logarithmic models, separated by condition and subtest. The most salient result was a clear advantage for linear models across all conditions and subtests, as there is no case where more participants' data within a condition and subtest was better fit by a log function than a linear function.

Table 4

Frequency of linear and log models by condition and subtest

Subtest	Log			Linear		
	NR	*CR*	*IR*	*NR*	*CR*	*IR*
1	3	7	1	24	20	25
2	6	11	5	21	16	21
3	8	10	5	19	17	21
4	10	11	3	17	16	23

Comparisons between conditions revealed a significant difference in the distribution of models for subtest 4 [$X^2(2) = 6.3$, $p = .04$], a trend towards significance for subtest 1 [$X^2(2) = 5.7$, $p = .06$], and no significant differences for subtest 2 or 3 [*subtest 2*: $X^2(2) = 3.6$, $p = .17$; *subtest 3*: $X^2(2) = 2.1$, $p > .2$]. Significant differences between conditions at subtest 4 likely reflect the relatively weak performance of both the *no ruler* and *congruent ruler* conditions in comparison to the *incongruent ruler* condition.

The application of qualitatively different logarithmic and linear models is supported by experimental evidence demonstrating a rapid shift between representative models following appropriate feedback during training (Opfer & Siegler, 2007). Alternatively, Barth and Paladino (2011) argue that same general pattern of development could emerge from refinement of a single quantitative model. Additionally, consistent nonlinear biases in purportedly linear estimates suggest alternative analysis. Specifically, the (inverted) s-curve of the cyclical power model (Hollands & Dyre, 2000; Hollands, Tanaka, & Dyre, 2002) often provides a better fit than linear models. Additionally, this model was developed to address tasks of proportional reasoning, which matches an interpretation of the number line task in which children are judging magnitudes as a proportion of the whole number line (Barth & Paladino, 2011; Barth, Slusser, Cohen, & Paladino, 2011).

Furthermore, as proportional reference points – e.g. the midpoint – are incorporated into the model multiple cycles of s-curves can be applied to fit equal divisions of the data. From the proportional reasoning account, a multiple-cycle model would indicate that the individual is comparing magnitudes to a partial segment of the entire number line. Because of the manipulation in this experiment, highlighting the

proportional relationships between landmarks, cyclical power models may offer increased discriminatory power between patterns of estimates produced by children in differing conditions.

To test their hypothesis regarding the superiority of the power model, Barth and Paladino (2011) test the fit of logarithmic, linear, one-cycle, and two-cycle power models directly. However, because of the difference in model complexity (i.e., number of parameters) an alternative criterion to minimum variance explained must be applied. One such measure, Akaike's information criterion (AIC), can be used to compare models with differing model complexity (Akaike, 1974, 1978). This measure applies a penalty for greater number of parameters. This statistic may be further corrected for a small number of observed values (Burnham & Anderson, 2002), by placing a stronger penalty on additional parameters. Here minimum corrected AIC is used to select between linear and logarithmic (two parameters), and cyclical models (one parameter). Table 5 displays the distribution of best-fitting models.

Table 5

Frequency of linear, log, and power regression model types by condition and subtest

Subtest	Log			Linear			One Cycle			Two Cycle		
	NR	*CR*	*IR*	*NR*	*CR*	*IR*	*NR*	*CR*	*IR*	*NR*	*CR*	*IR*
1	1	7	1	9	12	8	10	7	11	7	1	6
2	3	7	4	13	7	4	8	10	12	3	3	6
3	4	7	5	11	12	8	4	5	7	8	3	6
4	5	10	3	9	7	5	13	9	13	0	1	5

Comparisons between conditions revealed a significant difference in the distribution of models between conditions for subtest 1 [$X^2(6) = 14.1$, $p = .03$] and subtest 4 [$X^2(6) = 13.5$, $p = .04$], and no significant differences in subtest 2 or 3 [*subtest 2*: $X^2(6) = 9.4$, $p = .15$; *subtest 3*: $X^2(6) = 4.8$, $p > .2$]. Like the previous analysis, comparing only log and linear models, for each subtest children in the *congruent ruler* condition provided the highest frequency of logarithmic models, while children in the *incongruent ruler* condition provided the lowest frequency of logarithmic models. Unlike the log-linear analysis,

children in the *incongruent ruler* condition also provided the lowest frequency of linear models, implying that a large number of their estimation sets were re-interpreted in favor of a cyclical model.

While the cyclical power model was applied by Barth and Paladino (2011) as a direct alternative to the linear model (although they did account for logarithmic data with an additional parameter in the power model), Ebersbach, Luwel, Frick, Onghena, and Verschaffel (2008) developed an alternative specifically for the logarithmic model. Specifically Ebersbach et al. fit curved patterns of data with two connected linear segments. While the logarithmic model is more parsimonious than the segmented linear model (i.e., 2 parameters instead of 4, respectively), Ebersbach et al. argues that children simply maintain a normatively linear representation (i.e., slope approaching 1) of the subset of magnitudes that is highly familiar (e.g. 0-20 on a 0-100 scale), and perform nearly randomly for less familiar magnitudes. In a study of 5- to 9-year old children these researchers found that the break between linear segments reflected the upper limit of the children's counting ability.

While a break in linear segments could indicate transition from a familiar to unfamiliar subset of magnitudes, it could also reveal a more subtle transition between independent representations of partitions of the number line, promoted by the application of a landmark-based strategy. For example, a child could maintain two independent linear models of the numerical scale – one for the lower-half and one for the upper-half of data. To some extent this resembles the two-cycle power model; however, the segmented model does not assume that the patterns above and below the midpoint are symmetric.

While Ebersbach et al. (2008) used a four parameter model in which the location of the break point was estimated in the regression, the theoretical basis for the breakpoint in this experiment restricts its location to the midpoint (90). Therefore, instead of a four parameter model, here a three parameter model is used to derive parameters of the connected linear segments (intercept and slope of the first segment, slope of the second segment). Additionally, to address the possibility that these two linear segments were completely independent a four parameter model was included to derive disconnected linear segment (i.e., the intercept of the second segment is estimated by the model). To estimate these parameters, non-linear regression was performed in R. The distribution of models selected with the lowest AIC are displayed below in Table 8.

Table 6

Frequency of linear, log, and segmented regression model types by condition and subtest

Subtest	Log			Linear			Connected			Disconnected		
	NR	*CR*	*IR*	*NR*	*CR*	*IR*	*NR*	*CR*	*IR*	*NR*	*CR*	*IR*
1	2	4	1	17	14	18	4	9	5	4	0	2
2	6	10	3	14	10	13	4	3	4	3	4	6
3	8	9	3	15	14	13	4	4	4	0	0	6
4	8	8	3	14	14	16	4	3	6	1	2	1

Because the high cost of additional parameters the traditional log and linear models accounted for the majority of best fits. Comparisons between conditions revealed a significant difference in the distribution of models between conditions for subtest 3 [$X^2(6) = 15.4$, $p = .02$;], and no significant differences for the three other subtests [*subtest 1*: $X^2(6) = 8.8$, $p = .19$; *subtest 2*: $X^2(6) = 5.9$, $p > .2$; *subtest 4*: $X^2(6) = 4.4$, $p > .2$]. The significant difference between conditions at subtest 3, which did not emerge in either of the previous two analyses, appears to be driven by differences in distributions of disconnected linear segments models (6 for *IR,* 0 for *NR* and *CR*).

3.2.6. Estimation strategy

While the inclusion of partitioned models of estimates (i.e., two-cycle power, segmented linear) in the previous section was intended to test the hypothesis that children, following successful intervention, acted upon the number line by dividing it at its midpoint, this provide indirect evidence for differences in explicit strategy. Likewise, small but consistent differences in posttest trial durations between conditions (Figure 13) suggest that children approached the task differently for the differing conditions. An ANOVA of trial duration revealed a significant effect of condition [$F(2, 77) = 3.3$, $p = .04$, $\eta_p^2 = .08$], a significant effect of subtest [$F(3, 231) = 38.0$, $p < .001$, $\eta_p^2 = .33$], and no significant interaction between condition and subtest [$F(6, 231) = .05$, $p = 1.0$, $\eta_p^2 = .00$]. Bonferroni-adjusted post hoc analyses revealed a significant difference between children in the *no ruler* condition and children in the *congruent ruler* condition [$p < .05$].

To some extent faster trials by children in the *no ruler* condition may reflect the situational similarity between the learning task and the testing task. In contrast, children in the other two conditions transitioned from estimating with a ruler in the learning task to estimating without a ruler in the posttest. Therefore, the effort of constructing a new strategy may have resulted in longer durations. However, while one would expect that these differences to dissipate in the later subtests, this was not the case. The faster performance of children in the *no ruler* condition persisted throughout the entire posttest (i.e., no interaction of subtest and condition), suggesting that their learning experience affected some element of strategy.

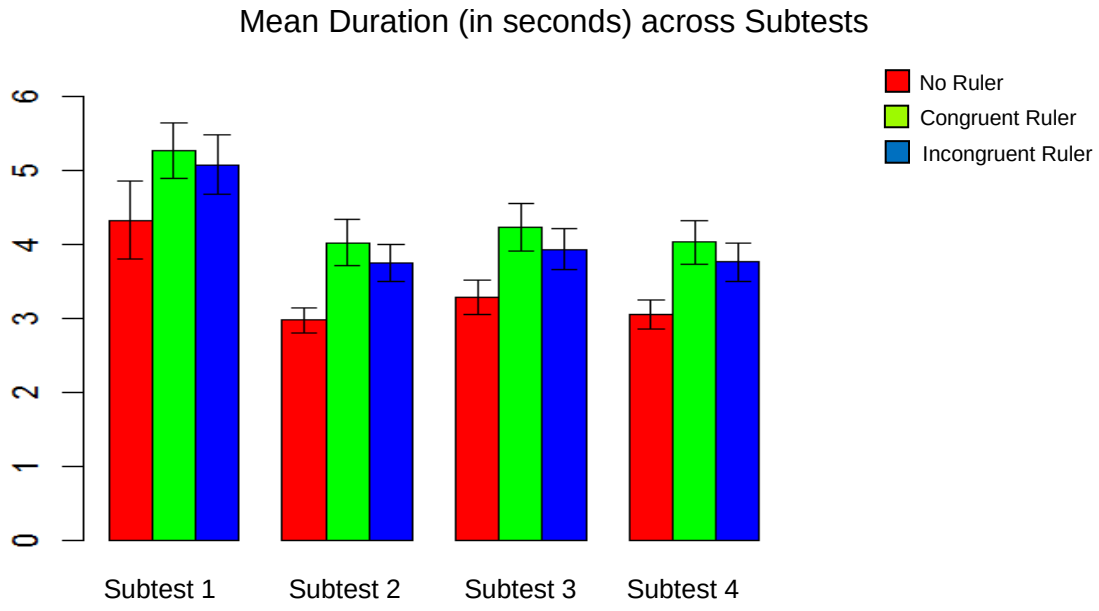


Figure 13. Mean duration (in seconds) for all participants across all four subtests.

In the case of ruler-based conditions the explicit recall of landmark information may have played a role in increased trial durations. For example, while estimating 81 a child might have recalled the location of 90 and used this landmark as a reference for his or her estimate. This strategy could be accomplished by simply fixating one's gaze on the landmark (which could be captured through eye-tracking), touching the screen at the location of the landmark (which could be captured through video or a touch-sensitive screen), or through the manipulation of the mouse cursor. In terms of the latter strategy, an emerging

technique of “mouse tracking” allows researchers to elucidate dynamic cognitive principles with commonplace resources (Farmer, Cargill, Hindy, Dale, & Spivey, 2007; Spivey & Dale, 2006).

Figure 14, below, demonstrates how a child’s use of numerical landmarks may appear in data. Specifically, as the child performed an estimation trial – moving the mouse along the number line – the current position of the mouse (in terms of magnitude of the given numerical scale) was captured at 20 millisecond intervals. In Figure 14, the left graph displays this mouse location information over time. Steep curves indicate rapid motion, while flat segments indicate slow or stopped motion. The right graph displays a density plot, indicating the proportion of time spent across magnitudes.

Example of a trial curve and associated density curve

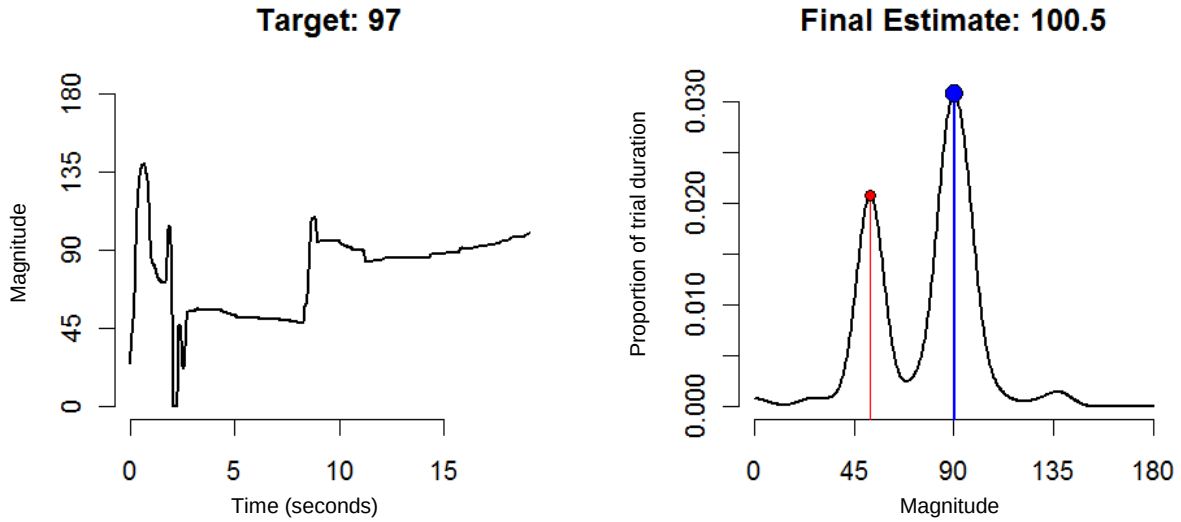


Figure 14. Left plot shows the position of the cursor – in terms of the associated magnitude on the number line – over time. In this case the subject was estimating the target 97 (in the training task), and produced a final estimate of 100.5. Long, flat stretches indicate magnitudes in which the subject either stopped or considerably slowed his or her mouse movements. On the right a density plot displays the magnitudes at which the participants lingered longest. In this case there were two clear peaks near 45 and 90. In an analysis of peaks near 90, the blue line represents a true positive, while a red line represents a false positive. Smaller peaks (such as one at 135) were disregarded.

In the case of Figure 14, the participant spent the greatest proportion of time near 45 and 90. This likely indicates that he or she was explicitly using numerical landmarks as reference points for the final estimate. In this case, the midpoint landmark of 90 well-affords estimation of the target 97. Furthermore, in the right graph the highlighting of the 90 peak reflects an analysis in which peaks near 90

(within 5% error, i.e., +/- 9 of 180) are marked. Yet, while the use of a midpoint reference is clear in this case, discerning between accidental stops near the midpoint and intentional, strategic stops presents a challenge. Peaks that are too shallow likely indicate random motion in the data. To some extent these "bumps" were removed by a smoothing function implicit in the density curve.

Furthermore, to buffer against detecting non-strategic variation in the data (i.e., false positives), only maxima whose difference from their closest minima exceeded 10% of the maximum height of the graph were included. For example, in Figure 14, the total height of the density curve (i.e., the highest maximum) was approximately .03. Therefore, to be detected, a peak had to exceed .003 difference from its nearest surrounding minima. The peak near 135 was approximately .002 higher than the minimum on its left, and therefore was not marked. Also, slowing down or stopping at the final estimate before clicking the mouse to end the trial was likely to result in a peak. If the final estimate was near 90, then a peak at 90 cannot be interpreted as indicative of a landmark strategy, but simply an artifact of approaching the final estimated magnitude. Therefore peaks within 5% of the final estimate were not detected.

This analysis was most likely to reveal differences between conditions during training – where the ruler marked landmarks visually, and many students explicitly navigated their mouse cursor to those locations. To analyze the use of landmarks, every trial was analyzed, according the procedure described above, to mark peaks near 90. Yet, while the use of a midpoint strategy was potentially applicable at all magnitudes, it was more likely at magnitudes near the center of the scale. Therefore, for each subject the number of trials with targets in the range of 67 to 113, containing a peak at 90, were counted. Each subject was then classified according to whether 90-peaks were present in at least one-third of relevant data (Table 4). As a comparison, participants with 90-peaks in one-third of trials with targets in the range of 0 – 23 and 167 – 180 were analyzed in parallel.

Table 7

Frequency of participants with 33% of trials containing peaks at 90 (Training)

Targets	Contained 90-peaks in 33% of data?	*No Ruler*	*Congruent Ruler*	*Incongruent Ruler*
67 – 113	Yes	5	9	13
	No	22	18	13
0 – 23, 167 – 180	Yes	1	4	4
	No	26	23	22

An analysis of the distributions contained in table 5 shows a (marginally) significant effect of condition [$X^2(2) = 5.9$, $p = .05$] for targets near 90 (67 – 113). A post-hoc, Bonferroni-adjusted, comparison between children in the *no ruler* condition and the *incongruent ruler* condition showed a trend towards a larger proportion of peaks at 90 for the *incongruent ruler* [$X^2(1) = 4.5$, $p = .10$]. No other direct comparisons neared significance. In the case of targets near 0 and 180 (0 – 23, 167 – 180, respectively), no significant effect of condition emerged [$X^2(2) = 2.3$, $p = .31$].

Although, at posttest, children no longer possessed an explicit reference for landmarks, i.e., the ruler, children may have applied the same strategy by imagining a ruler or by simply recalling the location of landmark values. Like training children were unlikely to use this strategy equally for all targets. Rather, target magnitudes of 81 and 106 were likely candidates for this strategy because of their proximity to 90. Furthermore, they were sufficiently distant from 90 to avoid confusion between final estimates and landmark strategies. Table 7 displays the number of participants in each subtest that had a peak at 90 in either the 81 or 106 trials.

Table 8

Frequency of participants with peaks at 90 for targets 81 or 106 (Training)

Subtest	Contained 90-peaks in 33% of data?	*No Ruler*	*Congruent Ruler*	*Incongruent Ruler*
1	Yes	4	7	8
	No	23	20	18
2	Yes	4	13	8
	No	23	14	18
3	Yes	4	10	10
	No	23	17	16
4	Yes	4	5	9
	No	23	22	17

An analysis of the distributions contained in Table 7 shows a significant effect of condition for subtest 2, only [$X^2(2) = 7.0$, $p = .03$]. A post-hoc, Bonferroni-adjusted comparison of children in the *no ruler* condition and the *incongruent ruler* condition shows a trend towards a significant difference during subtest 2 [$X^2(1) = 5.5$, $p = .06$]. While analyses of other subtests do not reveal significant results, in each case – as expected – children in the *no ruler* condition show the least number of stops at 90.

3.2.7. Verbal bisection probes

Finally, with the bisection questions that followed estimation, 3 *no ruler* participants, 17 *congruent ruler* participants, and 23 *incongruent ruler* participants correctly identified 90 as the midpoint of the 0-180 scale. Chi-squared tests revealed an overall effect of condition [$X^2(2) = 33.3$, $p < .001$]. Post-hoc tests, with Bonferroni-adjusted alpha levels, revealed a significant difference between participants in the *no ruler* condition and participants in both the *congruent ruler* condition [$X^2(2) = 13.4$, $p < .001$] and the *incongruent ruler* condition [$X^2(2) = 28.7$, $p < .001$]. For the second question, 1 *no ruler* participant, 7 *congruent ruler* participants, and 15 *incongruent ruler* participants correctly identified 45 as the midpoint of the 0-90 scale. Chi-squared tests revealed an overall effect of condition [$X^2(2) = 19.0$, $p < .001$]. Post-hoc test revealed a significant difference between participants in the *no ruler* condition and participants in the *incongruent ruler* condition, only [$X^2(2) = 15.9$, $p < .001$]. For the third question, 0 *no ruler*

participants, 5 *congruent ruler* participants, and 12 *incongruent ruler* participants correctly identified 135 as the midpoint of the 90-180 scale. Chi-squared tests revealed an overall effect of condition [$X^2(2) = 17.0$, $p < .001$]. Post-hoc test revealed a significant difference between participants in the *no ruler* condition and participants in the *incongruent ruler* condition, only [$X^2(2) = 13.6$, $p < .001$].

Table 9

Accurate verbal bisection distributions

Landmark value	Correct?	*No Ruler*	*Congruent Ruler*	*Incongruent Ruler*
90	Yes	3	17	23
	No	24	10	3
45	Yes	1	7	15
	No	26	20	11
135	Yes	0	5	12
	No	27	22	14

3.3. Discussion

3.3.1. Benefits of instruction with physical materials

Experiment 1 demonstrated that a physical representation of a numerical concept can either promote or interfere with learning, depending on the manner by which it is used by learners. In regards to the benefits of a physical representation, the results reveal an advantage for children in the *incongruent ruler* condition verses the children in the *no ruler* condition in both learning task performance and testing efficiency (as measured by total blocks to criterion). While the ruler was intended to guide the development of a specific representation of the number line, children without a ruler were less constrained in the development of their internal representation. While it is possible that some of the children in the *no ruler* condition spontaneously developed a quartile landmark-based strategy, this was rarely (if ever) the case.

As evidence, only 3 of 27 children in the *no ruler* condition, in comparison to 23 of 26 in the *incongruent ruler* condition, correctly identified 90 as the midpoint of 0 – 180. This clearly indicates that the true numerical midpoint played little role in *no ruler* strategy. However, in some cases *NR* children may have applied a spatial midpoint strategy with an incorrect estimate of the numerical midpoint. For example, five *NR* children misidentified the midpoint of 0 – 180 as 100. In at least one case, a child mentioned that the midpoint was 100 during estimation trials. While there was some, marginal, differences between children with and without rulers in analyses of trials curves (i.e., counting peaks at 90), the application of a 100-as-midpoint strategy could have masked differences between conditions. For example, a child who estimated 106 by stopping at the midpoint – which he or she believed to be 100 – would look strategically similar to a child who correctly understood the midpoint as 90.

Therefore, if a large number of children applied a spatial midpoint strategy with an incorrect numerical midpoint we would expect highly parallel results, with simply less accuracy. However, given the large variability in bisection probe answers (15 unique estimates, ranging from 17 to 160), overall lower trial durations, and a lack of anecdotal observation of related behavior, it is unlikely that many *NR* children took this approach. Yet, considering the large number of linear models fitting estimation data (vs. log models only, Table 4) for *NR* children, particularly at subtest 1 (24 of 27 sets of estimates were best fit by linear model), whatever representation these children developed was valuable in the posttest task.

The nature of the representation developed by children in the *no ruler* condition likely fit one of two possible descriptions (or some mix between them): either a holistic mapping between number and spatial magnitude that did not rely upon recall of specific spatial-numerical associations, or an accumulation of specific spatial-numerical associations that could be applied as a local reference. As an example of a general representation, when estimating 140 the child would simply recognize that this relatively large value should be situated at a relatively distant spatial location from zero.

On the other hand, with an accumulation of specific associations, a child might recall the location of a magnitude from specific feedback provided recalled from training (e.g. “140” – approximately 7 cm from the right endpoint). While it is unlikely that a child could recall all of the spatial locations of magnitudes presented during training, interpolation could be performed locally without resorting to a more general representation. For example, if the child recalled the position of 140 and 130, then he or she

could estimate nearby values (e.g. 135). This is, in some sense, a landmark-based strategy. However, unlike the quartile landmarks, which are defined by spatially-invariant structure, these landmarks were either arbitrary or bound to the local context. As such, a landmark strategy, devoid of spatially invariant landmarks, would introduce significant challenges to transfer. On the other hand, a more general association between number and magnitude, not reliant on specific recall, would promote high transfer.

In this experiment the use of multiple subtests, with varying spatial properties, was incorporated to test the flexibility of children's representations. As expected, a post-hoc comparison between the *incongruent ruler* and *no ruler* conditions did not produce a significant difference at subtest 1 – i.e., where either a specific associations or a general representation of magnitude could be applied successfully. However, by subtest 4 – where arbitrary relations between space and number could no longer be (easily) applied – differences between these conditions emerged.

Yet, while post-hoc tests revealed a different pattern of results for subtest 1 and subtest 4, this did not translate into a significant interaction between condition and subtest. An ANOVA of mean PAE, performed on a subset of data including only *NR* and *IR* participants and subtest 1 and 4, revealed a significant effect of condition [$F(1, 51) = 6.8$, $p = .01$, $\eta_p^2 = .12$] and block [$F(1, 51) = 30.7$, $p < .001$, $\eta_p^2 = .38$], but only a trend towards and interaction between condition and subtest [$F(1, 51) = 2.9$, $p = .09$, $\eta_p^2 = .06$]. Interpreted in isolation, this lack of interaction suggests that while children in the *no ruler* condition were, overall, less accurate, they were no more affected by the altered spatial layout of the later subtests than children in the *incongruent* ruler condition.

This use of the interaction term as a measure of far transfer assumes the "nearness" of the first subtest. Specifically, the initial subtest is presumed to be closely related to the training task, and therefore assesses the base accuracy of the representation developed in training. This should be particularly true in the case of the *no ruler* condition, where children are simply transitioning from training without a ruler to a nearly identical display during subtest 1. Considering this similarity between training and subtest 1, *NR* performance should have been expected to remain quite stable in this transition. Rather, the mean PAE of participants in the *no ruler* condition rose substantially from their last training block (Mean = 5% error, SD = 1.2%) to the first subtest (Mean = 9.5% error, SD = 3.5%).

To some extent this drop in performance was likely an artifact of regression to the mean (as the last block of training was the one in which they reached criterion); however, the transition from feedback trials to non-feedback trials may also have played a role in this decrease in accuracy. Furthermore, the abrupt change may have played a larger role for children in the *no ruler* condition than in the other conditions. In particular, given no structure for developing appropriate landmark values, *NR* children may have placed greater reliance upon previous trial feedback to estimate the current trial. For example, a child who had recently received feedback on the magnitude 133 could recall this information to apply as a landmark in estimating a subsequent magnitude of 140.

However, this strategy was not always equally accessible. Following each block a delay with an animation was presented to the participant. This animation lasted approximately 10 seconds, included an animation of a number of fish swimming, and a voice-over. Likely, by the time this animation had completed many children had forgotten their previous trial's feedback. For the remaining 7 of 8 trials, one trial followed immediately from the previous. If children in the *no ruler* condition were effectively using the feedback from the previous trial, their accuracy for the last 7 trials of a block would be expected to exceed their accuracy for the first trial of each block. In a comparison of mean PAEs of the first trial of blocks to the average of the last seven trials of the block (Figure 15), only participants in the *no ruler* condition showed a significant difference in accuracy [not assuming equal variances, $t(44.3) = 2.8$, $p < .05$].

This result suggests that the *NR* children's representation and strategy did not transfer fully intact to the first subtest. Rather, these children relied heavily upon on the recall of specific associations between number and space, which likely decayed rapidly. From this perspective, the lack of interaction between condition and subtest, but an overall effect of condition could be a consequence of a general inability to transfer from training to testing on the part of *NR* children. Anecdotally, children in all conditions often verbalized their recall of a specific magnitude location from training. While this occurred for landmark values for *IR* and *CR* children, this most often occurred for very small or large values (e.g. 5 and 178, respectively) for *NR* children.

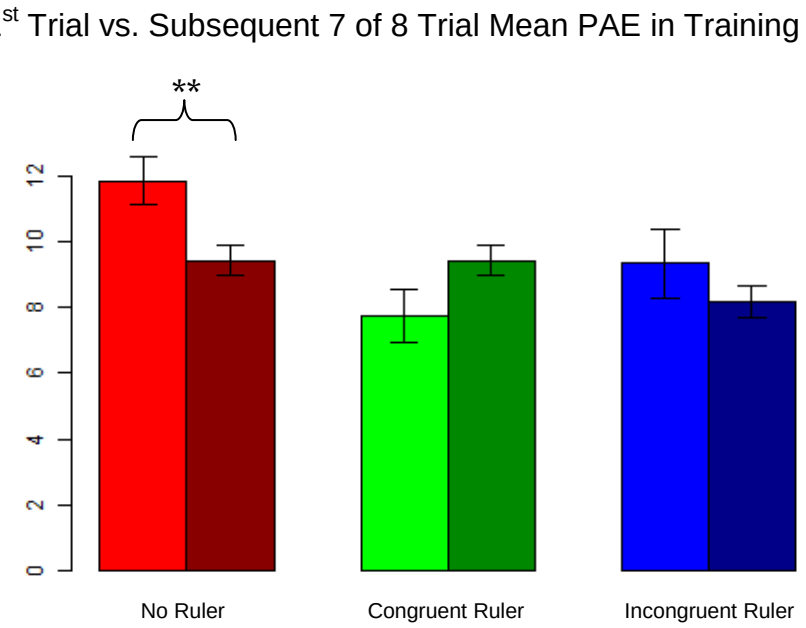


Figure 15. Mean Percent Absolute Error, comparing 1st trials in a training block to all seven others trials in a training block. 1st trial is shown on left, later trials are shown on right for each condition.

While it is clear that a well-developed quartile landmark representation benefitted children at posttest, another prediction was that the ruler would enable more efficient learning and higher performance during training. From an instructional point of view, this entails that appropriate guidance may not only help the student comprehend materials more deeply, but do so with less trial-and-error and failure in the learning process. While producing knowledge that is relevant beyond the learning context is the ultimate goal, doing so efficiently serves both motivational and time-management needs of the classroom. In the case of this study, those children in the *incongruent ruler* condition had a higher rate of success and engaged in a lower number of trials than their counterparts in the *no ruler* condition.

However, children in the *incongruent ruler* condition did spend, on average, more time on task than children in the *no ruler* condition (although, factoring in age caused this effect to dissipate). This difference was in part due to a large discrepancy in time of instruction, and in part due to implicit differences in strategies applied by children from these two conditions. Regarding the latter, children in the *incongruent ruler* condition required additional time to map values from their ruler to the target number line. While differences in time-on-task could be the underlying causal factor in posttest accuracy differences, given that *NR* children participated in, on average, more learning trials than *IR* children, it is difficult to interpret how time, itself, would be the source of greater knowledge in the *incongruent ruler*

condition. Most likely, given a more challenging criterion for training success (e.g. within 5% of the target magnitude), the differences between conditions in initial instruction would have played a less influential role than number of trials in determining time-on-task.

These results suggest that a spatial tool can play an important role in the development of a target representation. This clearly aligns with the proscriptions of constructivists that advocate for the use of concrete manipulatives in mathematics education (e.g. Sarama & Clements, 2009). However, as is expanded in the next section, when misapplied, concrete manipulatives can produce unintended and surprising deficits in a child's understanding.

3.3.1. Benefits of desirable difficulties

In this experiment the critical test of the desirable difficulties hypothesis was embodied in a comparison between children in the *congruent ruler* condition and children in either of the other two conditions. Specifically, the congruent ruler was designed to provide children with a highly accessible means of addressing the learning task, which required little trial-and-error or discovery. On the other hand, the elongated ruler was designed to provide a conceptual challenge that would need to be addressed by the learner, repeatedly, to achieve success in the learning task. The *no ruler* condition was included as a direct implementation of a discovery-oriented approach, where no instructional assistance was included. As expected children in the *congruent ruler* condition reached criterion in fewer trials and less time, and demonstrated less accuracy at posttest than the other two conditions.

Yet, while both the *incongruent ruler* and the *no ruler* condition were conclusively more difficult than the *congruent ruler* condition, the nature of these difficulties was very different. For the *no ruler* condition the challenge for most learners was (presumably) developing a novel representation and estimation strategy with little relevant prior knowledge or direction. On the other hand, for children in the *incongruent ruler* condition, the representation and strategy were implied by the ruler. For these children, the challenge was simply a matter of interpreting and internalizing the given representation. While there may be a benefit to developing one's own representation for a concept from scratch, given the advantages of the *incongruent ruler* condition it seems that a more directed form of conceptual difficulty is sufficient.

Given the advantage of combined *NR and IR* conditions to the *CR* condition, and the advantage of *IR* to *NR*, it may be deduced that the conceptual challenge of the *incongruent ruler* condition benefitted participants for reasons beyond simple exposure to the landmark magnitudes. Yet, the large differences between children in the *congruent* ruler condition and *incongruent ruler* condition on time-on-task and number of trials completed, provides grounds for alternative explanations for differences in posttest performance. Specifically, one may argue that the children in the *incongruent ruler* condition were successful simply because they had more opportunities to view the landmark magnitudes. Although children in the *congruent ruler* condition may have demonstrated rapid success in the task, the process of internalizing features of the ruler simply may have required more time and estimation trials.

Furthermore, the limited difficulty of the *congruent ruler* condition could have made available higher-level cognitive resources that would have benefited these children. By training to criterion children lost the opportunity to learn in this cognitively flexible state. From this perspective, the benefit of learning with highly intuitive materials may not be evident during the initial learning phase, but in a later phase when the learner is no longer struggling to form an initial conceptual representation. While, previous research has demonstrated limited success in these cases of “over-learning” (Bjork, 1994), I tested this assertion in the following experiment.

4. Experiment 2

As discussed above, significant differences in the time-to-criterion between the *congruent ruler* condition and the *incongruent ruler* condition undermines the interpretation that desirable difficulties are a necessary component of successful learning. In this experiment I address this by equating children by time-on-task. I chose to equate conditions on time, rather than number of trials, because it is likely that children in the *congruent ruler* condition, with a less challenging task, would complete more trials, in the same amount of time, as children in the *incongruent ruler* condition. Thus, children in the *incongruent ruler* condition should, on average, perform the learning task in equal time and fewer trials than children in the *congruent ruler* condition – eliminating a potential alternative explanation.

In experiment 1 the older children (third grade and fourth grade students) demonstrated less difference between conditions. For example, Appendix C-2 shows median responses of older children that appear universally linear. On the other hand, Appendix C-3 shows median responses for younger children that, in some case, appear to be approaching a logarithmic function; particularly with children in the *congruent ruler* condition. Therefore to achieve more variability in posttest measures this study only included second grade students.

Finally, while experiment 1 specifically addressed the issue of transfer over spatially-transformed posttest stimuli, other forms of transfer may also be promoted by the *incongruent ruler*. Specifically, because a quartile-based representation can be applied to any numerical scale by numerical division, children may spontaneously transfer their strategy to a novel numerical scale. Therefore, to test transfer across numerical scales, a subtest with a novel numerical scale was introduced.

4.1 Method

4.1.1. Participants

Participants included a total of 30 second grade students. Children were recruited from two afterschool programs in the New York City area serving primarily low-income Hispanic and African American populations. The *congruent ruler* condition included 15 children (M = 8.7 years, SD = .86, 48% female, 93% Hispanic, 7% African American) and the *incongruent ruler* condition included 15 children (M = 8.5 years, SD = .72, 58% female, 92% Hispanic, 8% African American).

4.1.2. Experimental design

The overall design of this experiment was the same as experiment 1 except that children were trained for 15 minutes (plus additional time to complete a block), instead of to criterion.

4.1.3 Materials and procedure

4.1.3.1. Standardized measures. To ensure that children from each condition had similar levels of mathematical achievement the Woodcock Johnson III Calculation and Mathematical Fluency subtests were administered in small groups of mixed-condition students in a quiet room.

4.1.3.2. Number line estimation training game. Training was administered one-to-one in either a private room, or in a private area of a large room. The child was placed at a desk with a computer, while the experimenter sat to the child's side to provide assistance. The training software and physical materials (rulers) were identical to those used in experiment 1.

However, unlike experiment 1, after completing the short animated instructional sequence – introducing children to the context and goals of the game – the administrator began a fifteen minute timer. The children then received the same condition-specific instruction as in experiment 1, before proceeding to estimation trials. After 15 minutes had expired the child was allowed to complete his or her current block of estimation trials before the learning task was halted by the administrator.

4.1.3.3. Number line estimation posttest. Like experiment 1 the first subtest of experiment 2 contained a number line that was equivalent in length and orientation to the training number line. However, to foster a more abrupt change, and minimize the duration of the study for the selected population of younger children, only the vertical (half-scale) number line was used to test transfer over spatial transformation (i.e., subtest 4 from experiment 1).

To extend the results of experiment 1, a final subtest, spatially equivalent to the first, tested children's ability to estimate on a 0 – 90 scale. In this case, the children were explicitly alerted to the new scale, and reminded throughout the subtest (i.e., children were told, "where is 24, between 0 and 90?"). The set of target magnitudes was constructed by dividing each value from the previous set of in half, and rounding up (3, 8, 16, 18, 23, 25, 29, 35, 42, 45, 47, 53, 60, 66, 68, 70, 78, 81, and 89).

The follow-up bisection questions remained the same from experiment 1 to experiment 2.

4.2. Results

4.2.1. Standardized measures.

Participants in the *congruent ruler* condition received a mean standardized score, grade-normed, of 100.9 on Math Fluency (SD = 15.2) and 101.2 on Calculation (SD = 8.4). Participants in the *incongruent ruler* condition received a mean standardized score, grade-normed, of 94.1 (SD = 12.2) on Math Fluency and 101.5 on Calculation (SD = 9.4). T-tests results show no significant difference between groups for either subtest [Math Fluency: t(28) = 1.3, p = .19; Calculation: t(28) = -.10, p = .92].

4.2.2. Number line estimation training game.

In this experiment children were equated on total duration instead of criterion performance (or number of blocks). However, to ensure that all participants completed the blocks that they began, total duration may have differed minimally between participants. Furthermore, because of differences in instruction and participant strategies, affecting trial durations, the number of total blocks may have differed between groups. However, two-tailed t-test revealed no significant difference between condition in either total duration [t(28) = -.53, p = .60], or total blocks [t(28) = 1.35, p = .19].

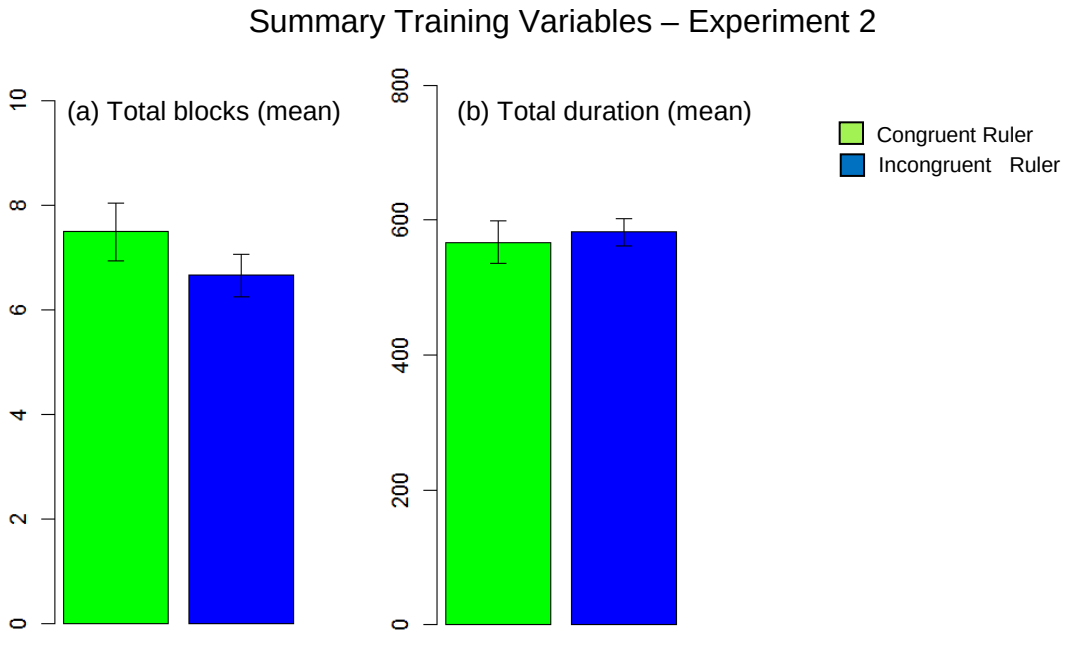


Figure 16. Summary training variables: (a) The number of blocks needed to reach criterion. (b) The total duration for the entire training session (in seconds).

Unlike experiment 1, in which some participants completed the training within a single block, in experiment 2 all children completed at least four blocks [*congruent ruler:* min = 5, max = 11; *incongruent ruler:* min = 4, max = 9]. Thus, while in experiment 1 blocks 1 and 2 were analyzed separately (because of differing number of participants), in experiment 2 the first four blocks of training were analyzed with a repeated-measures ANOVA (similarly to post-subtests). For example, Figure 17 displays the mean PAE of the first four blocks of training in experiment 2, as well as the overall average. An ANOVA of mean PAE reveals a significant effect of block [$F(3, 84) = 8.8$, $p < .001$, $\eta_p^2 = .24$], but no significant effect of condition [$F(1, 28) = .37$, $p = .55$, $\eta_p^2 = .01$] or interaction between block and condition [$F(3, 84) = .83$, $p = .48$, $\eta_p^2 = .03$]. Thus, while children were improving on successive blocks, condition did not appear to play a significant role in the magnitude of errors during training.

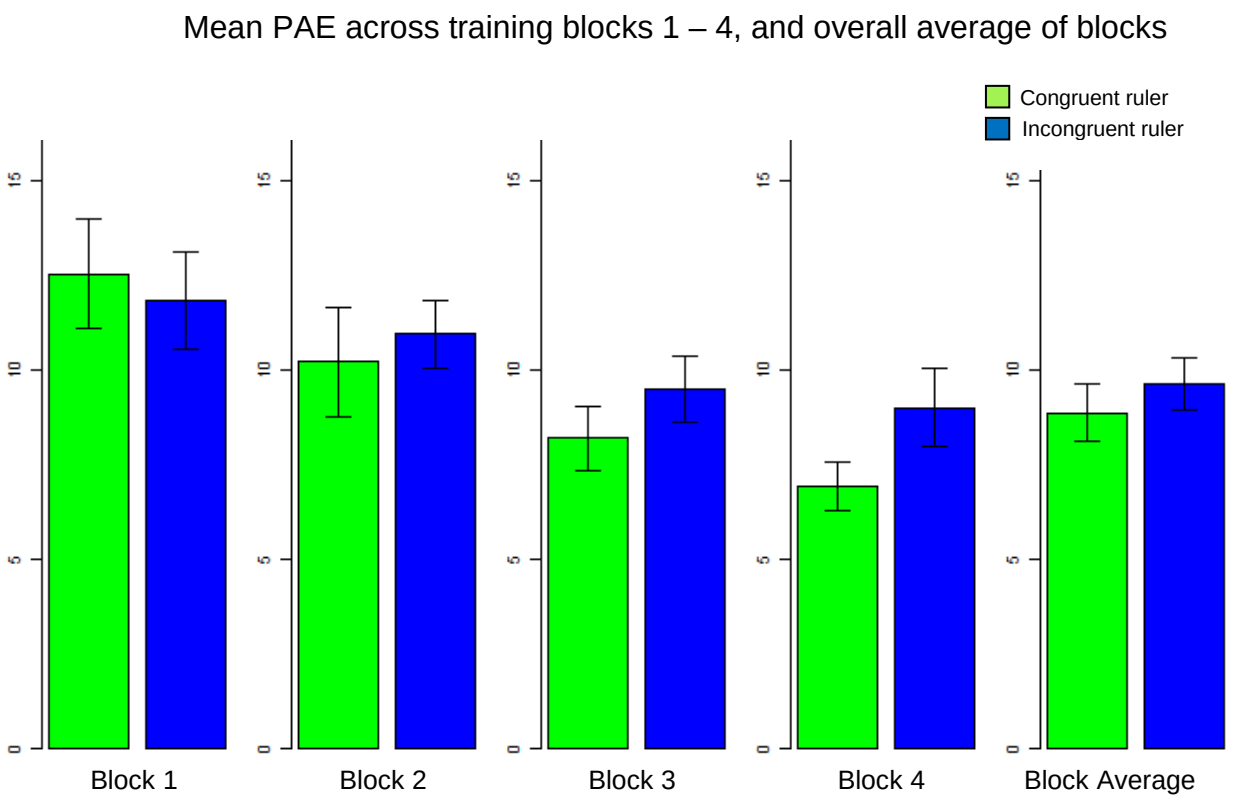


Figure 17. The mean PAE of participants in the first four blocks in training. The final graph shows the mean of each child's overall mean PAE.

A similar pattern emerged for the total durations of blocks. As displayed in Figure 18, children appeared to complete later blocks faster than earlier blocks. Somewhat surprisingly, the average change from block 1 to block 2 did not appear as dramatic as in experiment 1. An ANOVA on total block duration revealed a significant effect of block [$F(3, 84) = 75.1$, $p < .001$, $\eta_p^2 = .73$], but no significant effect of condition [$F(1, 28) = .87$, $p = .36$, $\eta_p^2 = .03$] or interaction between block and condition [$F(3, 84) = 1.27$, $p = .29$, $\eta_p^2 = .04$].

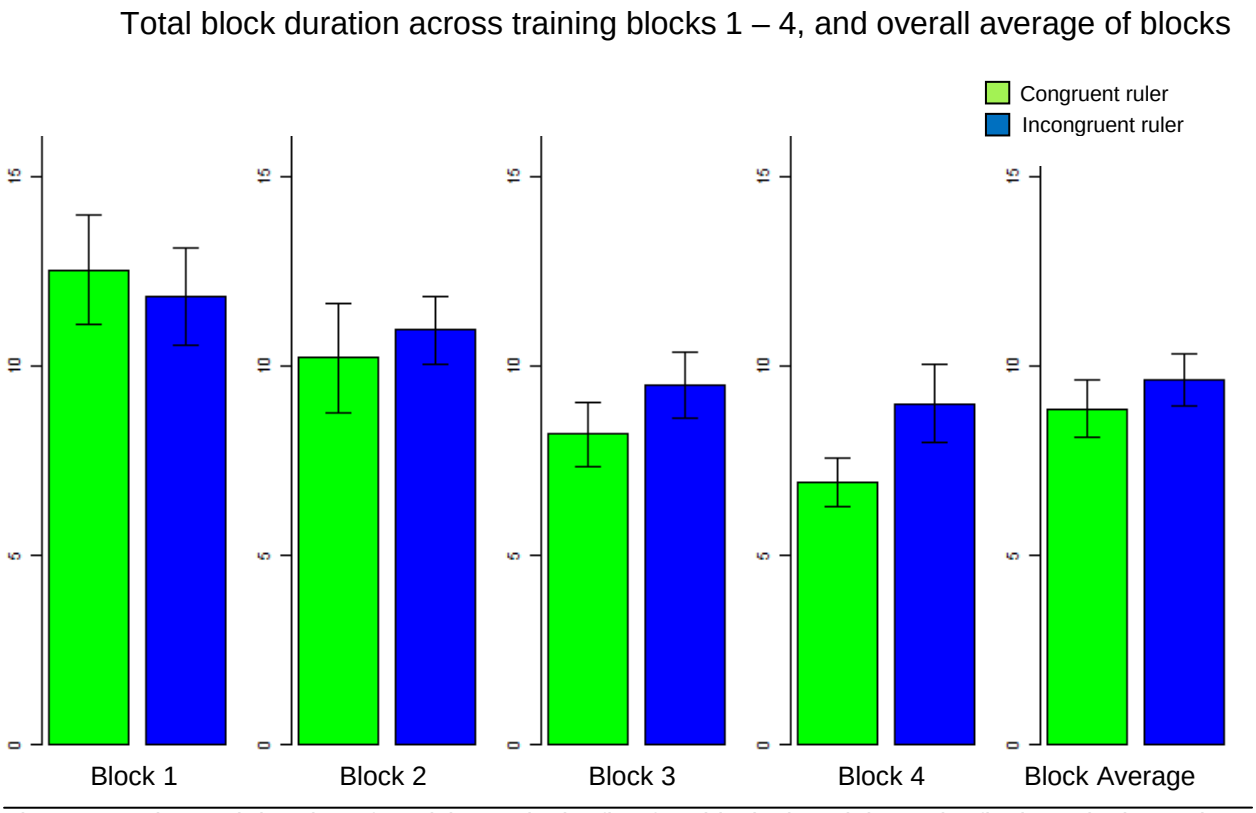


Figure 18. The total duration of participants in the first four blocks in training. The final graph shows the overall mean of all blocks total duration.

Finally, given the lack of differences between condition for both duration and PAE, I conducted an additional analysis of the total number of "correct" estimates per block (i.e., the number of "fish caught", Figure 19). While this measure is closely related to mean PAE, it represents the direct objective of the children. An ANOVA on total correct per block revealed a significant effect of block [$F(3, 84) = 4.4$, $p = .007$, $\eta_p^2 = .14$], a trend towards a significant effect of condition [$F(1, 28) = 3.15$, $p = .09$, $\eta_p^2 = .10$], and no

significant interaction between block and condition [$F(3, 84) = .73$, $p = .54$, $\eta_p^2 = .03$]. Therefore, in this case, an effect of condition approaches significance.

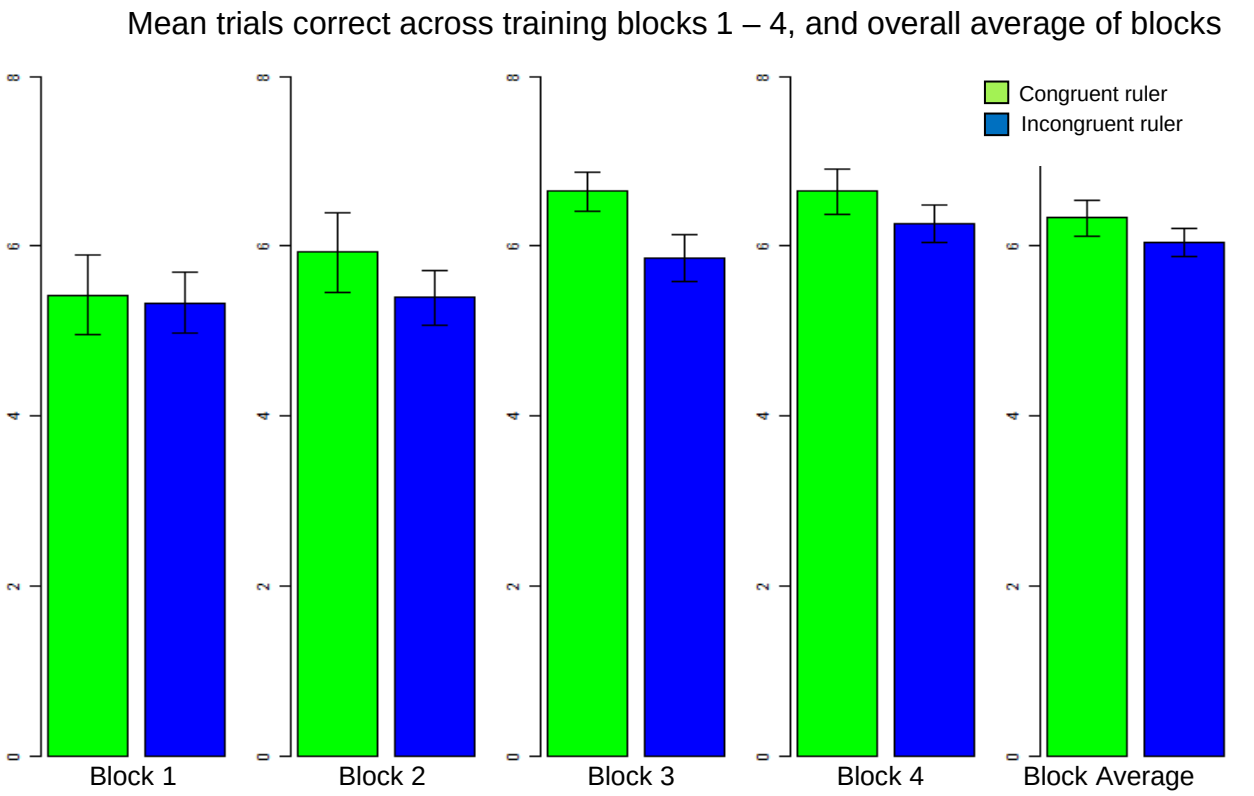


Figure 19. The mean trials correct of participants in the first four blocks in training. The final graph shows the mean of each child's overall average trials correct.

4.2.3. Posttest accuracy measures.

Like experiment 1, the median estimates over actual magnitudes did not show qualitative differences in curve type (i.e., all are linear, as displayed in Appendix I). Therefore, analyses of accuracy measures – including mean PAE, the variance explained by the linear fit (linearity), and the slope of the linear fit – were applied to demonstrate differences between experimental conditions. To avoid confusion with subtests from experiment 1, the first subtest of experiment 2 was re-labeled "equivalent" to indicate its physical similarity to the training number line, the second subtest was labeled "spatial transfer" to indicate that it was spatially dissimilar (i.e., rotated and scaled) to the training number line, and the third subtest was labeled "numerical transfer" to indicate that it was spatially equivalent but numerically distinct from the training number line.

Figure 20, below, displays the mean PAE's of experiment 2 participants across each subtest. An ANOVA of mean PAE revealed a significant effect of condition [$F(1, 28) = 8.8$, $p = .006$, $\eta_p^2 = .24$], a significant effect of subtest [$F(2, 56) = 4.9$, $p = .01$, $\eta_p^2 = .15$], but no significant interaction between condition and subtest [$F(2, 56) = .32$, $p = .73$, $\eta_p^2 = .01$]. T-tests comparisons between conditions at each subtest revealed a significant difference between conditions in the *equivalent* subtest [$t(28) = 2.9$, $p = .007$], a a significant difference between conditions in the *spatial transfer* subtest [$t(28) = 2.5$, $p = .02$], and a trend towards a statistically significant difference between conditions in the *numerical transfer* subtest [$t(28) = 1.7$, $p = .09$].

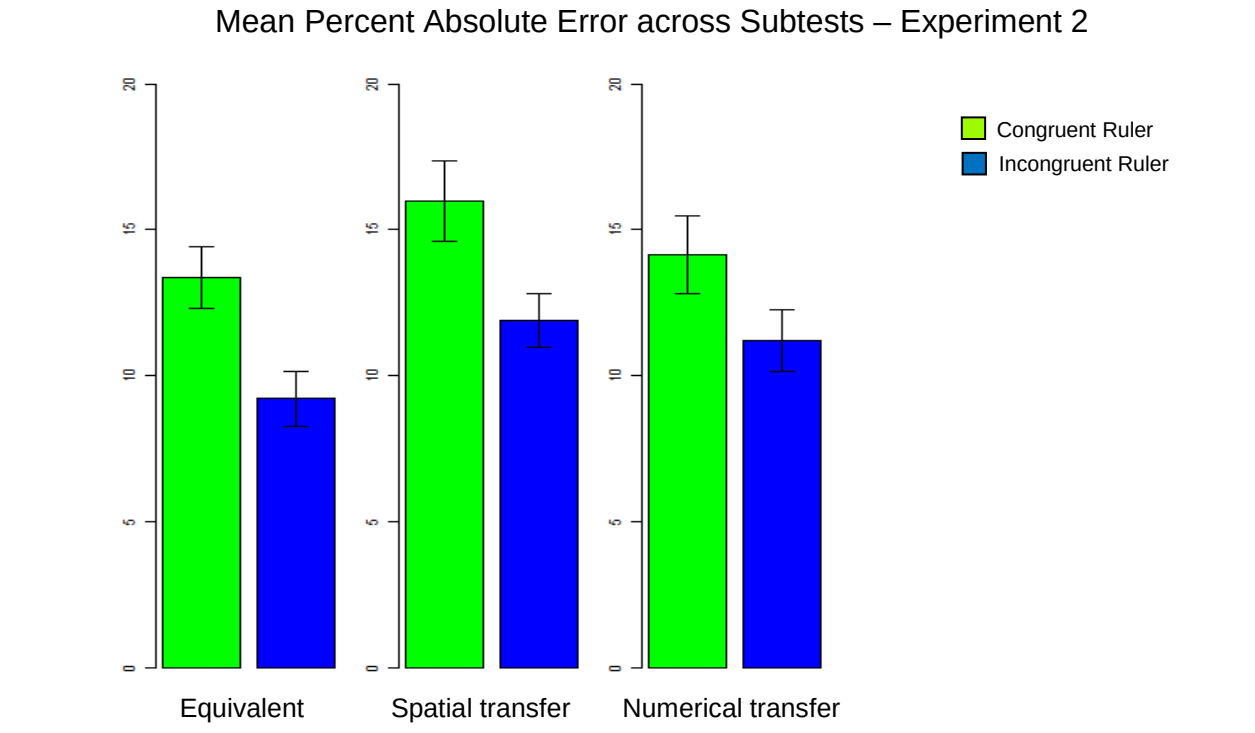


Figure 20. Mean percent absolute error for all participants, across three subtests.

Likewise, an analysis of linearity (Figure 21) demonstrated similar effects of condition to those described for mean PAE, but perhaps reduced differences between subtests. An ANOVA of linearity revealed a significant effect of condition [$F(1, 28) = 5.7$, $p = .02$, $\eta_p^2 = .17$], no significant effect of subtest [$F(2, 56) = 2.3$, $p = .11$, $\eta_p^2 = .08$], and no significant interaction between condition and subtest [$F(2, 56) = 1.8$, $p = .17$, $\eta_p^2 = .06$]. Analysis of simple effects at each subtest revealed a a significant difference

between conditions in the *equivalent* subtest [t(28) = 2.3, p = .03], a significant difference in the *spatial transfer* subtest [t(28) = 2.7, p = .01], and no significant difference in the *numerical transfer* subtest [t(28) = 1.1, p > .2].

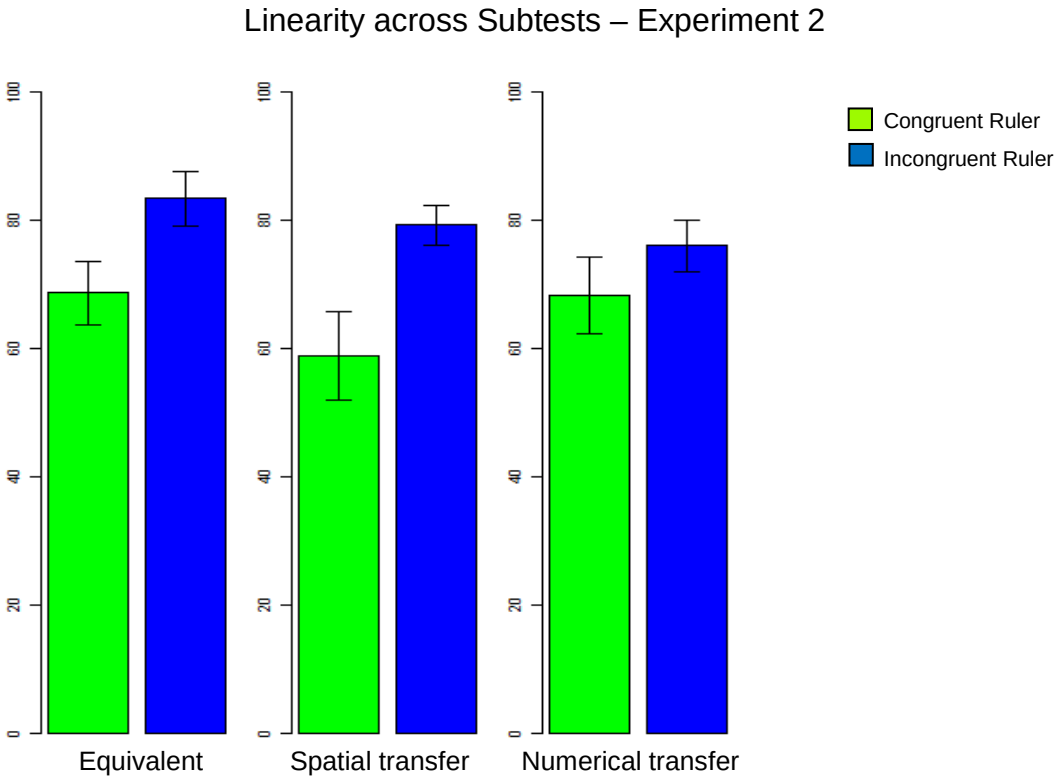


Figure 21. Mean linearity for all participants across all three subtests.

Finally, Figure 22, shown below, displays slope means. An ANOVA of slope revealed a trend towards significant effect of subtest [$F(1, 28) = 3.4$, $p = .08$, $\eta_p^2 = .11$], a significant effect of condition [$F(2, 56) = 8.0$, $p = .001$, $\eta_p^2 = .22$], and no significant interaction between condition and subtest [$F(2, 56) = 1.2$, $p = .32$, $\eta_p^2 = .04$]. Analysis of simple effects at each subtest revealed a significant difference between conditions in the *equivalent* subtest [t(28) = 2.3, p = .03], a trend towards a significant difference in the *spatial transfer* subtest [t(28) = 2.0, p = .06], and no statistical difference at the *numerical transfer* subtest [t(28) = .45, p > .2].

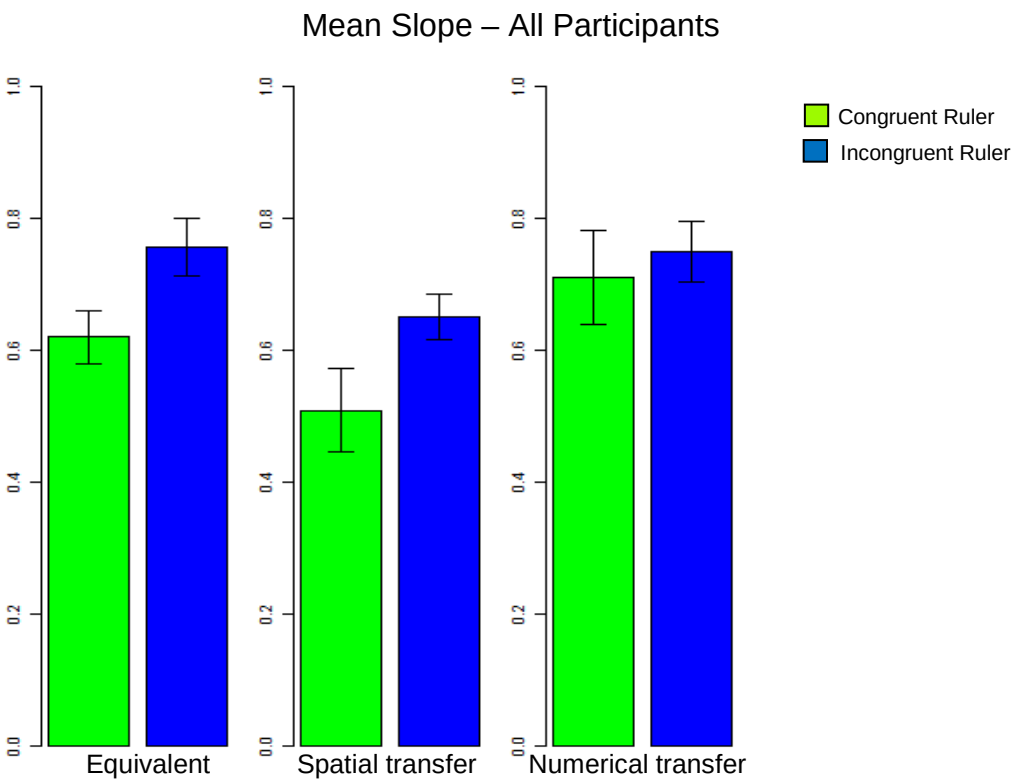


Figure 22. Mean slope for all participants across all four tests.

In experiment 1, the tails of the target range (5, 178) produced a great deal of leverage on the slope, and were removed for a second analysis. Applying this procedure here only affected the overall patterns marginally. Specifically, An ANOVA of slope (without 5 or 178) revealed a trend towards significant effect of subtest [$F(1, 28) = 3.4$, $p = .06$, $\eta_p^2 = .12$], a significant effect of condition [$F(2, 56) = 6.2$, $p = .004$, $\eta_p^2 = .18$], and no significant interaction between condition and subtest [$F(2, 56) = 1.3$, $p = .27$, $\eta_p^2 = .05$]. Analysis of simple effects at each subtest revealed a significant difference between conditions in the *equivalent* subtest [$t(28) = 2.6$, $p = .02$] and the *spatial transfer* subtest [$t(28) = 2.3$, $p = .03$], and no statistical difference at the *numerical transfer* subtest [$t(28) = .38$, $p > .2$].

4.2.4. Additional posttest estimation analyses

In experiment 1 each participant's data was fit by a series of models to compare distributions between conditions. Results revealed that distributions differed between conditions in a limited number of cases. In this experiment similar results would be expected; however, with only 15 participants per

condition, chi-squared tests of independence did not reveal distributional differences between conditions. Associated tables are displayed in Appendix K. Additionally, each set of participant's subtest data, and best fitting model are displayed in Appendix J.

In addition to model analysis, experiment 1 analyzed the estimation strategy of participants. As described above, there was no statistical difference in trial durations during training. Likewise, an ANOVA of trial duration during posttest revealed no significant effect of condition [$F(1, 28) < 1$] or interaction of condition and block [$F(1, 28) = 1.6$, $p > .2$]. Because differences in trial curve analyses appeared to be driven by the *no ruler* condition, which was not used here, this analysis is not presented here. A preliminary analysis revealed no significant differences between conditions.

***4.2.5. Verbal bisection probes*.**

Finally, in experiment 1, children in either ruler condition were more likely to correctly answer the bisection questions than children in the *no ruler* condition. However, a small difference between *IR and* CR conditions trended towards significance, in favor of *IR*. Yet, this effect may have been due to the comparatively greater exposure to the task that the *IR* children experienced. Table 6 shows that this is a likely explanation for the difference for experiment 1. In experiment 2, these groups only differed non-significantly on estimating the midpoint of 90 [$\chi^2(1) = .17$, $p = .68$], and not at all for estimations of 45 and 135 – which was poor, overall.

Table 6

Accurate verbal bisection distributions

Landmark value	Correct?	*Congruent Ruler*	*Incongruent Ruler*
90	Yes	10	12
	No	5	3
45	Yes	3	3
	No	12	12
135	Yes	1	1
	No	14	14

4.3. Discussion

This experiment confirms that the differences between the *congruent ruler* condition and the *incongruent ruler* condition in estimation accuracy, seen in experiment 1, were not due to differences in time-on-task, or number of trials. Clearly, these effects were due to the nature of the manipulation. Specifically, by impeding the physical coordination of the ruler to the on-screen number line, the incongruent ruler afforded development of a more appropriate and flexible cognitive process.

Surprisingly, however, there was little noticeable difference between treatment groups in the training performance. This appears to contradict an a priori assumption that the *congruent ruler would* produce higher learning task performance. The only evidence of this performance advantage for the *congruent ruler* was a trend in the average number of trials correct across the first four blocks. Yet, in all of the measures, differences fell in the expected direction (in favor of the *congruent ruler*). Thus, while it is tempting to claim that the additional challenge of the *incongruent ruler* actually came at little or no cost to learning efficiency, perhaps with more participants or a slightly more challenging task, a difference would have emerged.

Also, unlike experiment 1, no differences between conditions in best-fitting models were demonstrated here. Similarly, the conditions did not differ significantly in the measure of posttest slopes of best-fitting linear models. Therefore, it is possible that the additional time-on-task for *CR* (or inversely, the reduction in time-on-task for *IR*), contributed to greater similarity in the overall shape of the data. Most likely, with greater time on task, children in the *CR* condition had the opportunity to refine their representation of values near 0 and 180, which produce significant leverage on model types and slope, and transferred this representation to the posttest. On the contrary, magnitudes near the center, whose location was dependent on the ruler, were less likely to be recalled at posttest, therefore contributing to higher levels of overall error and reduced linearity compared to those who internalized magnitudes via the *incongruent ruler*.

While the analyses revealed general differences between conditions for most measures, analyses of the third subtest (numerical transfer), were marginal or inconclusive. Unlike the spatial transfer task, where the relative position (proportional distance) of any given magnitude on the number line remained

the same as the training task, here the magnitude needed to be reinterpreted in terms of a smaller scale. Training with the incongruent ruler was intended to support this activity by highlighting the relationship between 45 and 90. However, it was not the case that many children explicitly discovered this relationship. Only 3 of 15 children in each condition correctly estimated 45 as the midpoint of the 0 – 90 scale during the verbal bisection probes.

Yet, even with a non-midpoint-based strategy, children could accurately estimate values on this new, numerically-shortened number line with a more general representation of magnitude. Unfortunately, it seems, specific spatial-numerical associations appeared to be highly influential and produced some unexpected results. While performance on the training-equivalent subtest and the spatial transfer subtest were highly related [$r = .71$, $p < .001$], the numerical transfer subtest was less closely related to performance on the training-equivalent subtest [$r = .56$, $p = .001$], and nearly unrelated to the spatial-transfer subtest [$r = .33$, $p = .08$]. In other words, children's performance on subtest 3 often did not depend on performance on subtest 2 (which came as a surprise to the experimenter).

Most unpredictably, unlike experiment 1, or any known prior study of the number line estimation paradigm, a large number of participants produced estimates that resembled an exponential curve on this final subtest. A re-analysis of best-fitting models revealed that of the 30 participants, 13 sets of estimates on the numerical transfer subtest were best fit by an exponential function (8 linear, 9 logarithmic). The experimenter observed that this effect was likely caused by children's non-adaptive use of recalled spatial-numerical associations. For example, having just completed estimation on the 0 – 180 number line, a child might have recalled that the value of 29 was very close to the left-most endpoint, and estimated this magnitude at a similar location on the 0 – 90 number line, not taking into account its new proportional relationship to the whole number line.

Yet, if specific recall of spatial-numerical associations represented the entirety of the child's strategy his or her data would best be fit by a linear function with a slope of approximately .5. However, in addition to this specific recall of magnitude locations, children also appeared to use a more general association between number and space – particularly at the higher end of the scale. Higher values, such as 81, were used to fill the space between the midpoint and the right endpoint of the number line, instead of being placed in the same location as they would have been on the previous, 0 – 180 number line.

Because children used specific associations to crowd a large number of estimates in the first half of the number line and used a more general association to estimate a relatively few number of higher magnitudes at the far end of the number line, a characteristically exponential pattern emerged in nearly half of the cases. The reason why children applied specific memories for lower magnitudes, and a more general association for higher values – and not, for example, vice-versa – is unclear. However, it may be the case that feedback for smaller magnitudes was more salient during training because of the greater discrepancy with the intuitive, logarithmic model.

Similarly, several of the children who produced logarithmic patterns of estimates in the training-equivalent and spatial transfer subtests produced linear estimates at the numerical transfer subtest. As explained by Ebersbach et al. (2008), another interpretation of logarithmic results is a split between two linear segments. The first segment, most likely, rises quickly, while the second is relatively shallow. For these children on the first two subtests, the subset of data from 0 – 90 on the 0 – 180 scale was essentially linear. By constraining these children to only estimates of 0 – 90 on the numerical transfer subtest, only this linear component of their overall logarithmic model was required. For example, while a child may have incorrectly estimated the location of 45 on the 0 – 180 number line near the midpoint, this estimate would be accurate for the 0 – 90 number line. Furthermore, inaccurate estimation of larger magnitudes, which may have been underestimated by a logarithmic model (e.g. 155), played no role in estimation at the 0 – 90 scale.

In these cases, specific spatial-numerical associations appeared to play a more determinant role in estimation than a general sense of magnitude. Children who formed strong memories for previous trial feedback misapplied this knowledge on the new numerical scale, across both conditions. While the *incongruent ruler*, perhaps assisted in the internalization of a specific set of spatial-numerical associations, it is not clear that the children viewed these landmarks, other than the midpoint (90), in terms of their spatial and numerical significance. If the children had understood that 45 was marked because of its role as the midpoint between 0 and 90, they could have applied this knowledge during the numerical-transfer subtest, or conveyed this knowledge during the verbal bisection probes. Rather, many children appeared to associate the value of 45 with a static distance from zero. While children could

scale this distance on the spatially transformed number line, they could not reinterpret the meaning of 45 on the 0 – 90 number line.

To some extent the trouble with 45 and 135 may be inherent in the novelty of these particular numbers as important markers. Children likely expect multiples of 10 to have numerical and spatial significance in the range of values. Including two numbers that do not have a zero in the ones place may have promoted some confusion. Additional study with other scales (e.g. 0 – 200) would address this concern. Beyond the particular numbers chosen for this study, it is also clear that some improvements can be made in the tool, itself – which I discuss in the following chapter.

5. General Discussion

5.1. Development of a landmark-based strategy

In previous number line research transfer was usually studied in terms of different numerical scales (e.g. 0 – 1000) or different numerical tasks (e.g. arithmetic computation). Here, I primarily take the approach of keeping the task and numerical scale the same and altering the spatial display of the number line (except for the last subtest of experiment 2). Using this approach I analyze the extent to which numerical knowledge is bound to a specific visual display, and the extent to which it is flexible. In nearly all results, as the spatial display was transformed from the original training display, performance suffered (i.e., large effects of subtest across all measures).

To some extent this reduction in accuracy was reflected in a small shift in the distribution of linear and logarithmic models of estimates from 69 linear and 11 logarithmic at subtest 1 to 56 linear and 24 logarithmic at subtest 4 [$\chi^2(1) = 5.3$, $p = .02$]. Yet, even within the *congruent ruler* condition, more than half of the participants (16 of 27) produced linear estimates during the final subtest. Because a pretest was not included into the design of this study I cannot be sure that these 16 participants did not enter the experiment with intact linear representations of the 0 – 180 number line. However, I suspect, based on previous work on degree estimation with a similar population (Vitale, Black, Carson, & Chang, 2010) that children in all three conditions learned from the training task.

Yet, if a majority of children were able to develop linear representations by posttest, why did significant differences between conditions emerged in the analysis of mean errors and linearity? To some extent this may be related to the sensitivity of the measures. Classifying participants' data in terms of a small set of possible categories inevitably leads to loss of information. On the other hand the mean PAE and linearity measures (and to some extent slope) are affected by each data point.

Beyond the strength of the measures, it is possible that there are two processes that occur in the development of numerical representation: an initial qualitative shift, followed by local refinement of the representation. As Opfer and Siegler (2007) demonstrated, this shift may occur rapidly, in as little as one or two feedback trials. Even in Barth and Paladino's (2011) alternative cyclical power model, a rapid shift may occur between an unbounded power function (i.e., a power function that does not cycle within the

bounds of the numerical scale, producing logarithmic-looking data) to a bounded power function (whose cycle terminates precisely at the end of the scale). In this experiment children may have produced linear estimates by learning how to estimate magnitudes nearest to the endpoints first – which is suggested by the increased leverage on slope of the target 5 and 178. In all conditions, children learned to estimate very small and very large numbers with reference to displayed endpoint values which persisted through the posttest. By developing spatial-numerical mappings from the outside-in, the characteristically logarithmic biases in estimates may have been lost rapidly (i.e., the linear and logarithmic curves are most similar near the center of the scale).

In terms of the latter, refining process, specific mappings between space and number were likely developed during training for mid-range magnitudes. The applicability of these mappings in spatial-transfer subtests was likely a function of the context-specificity of the cues that produced them. In the case of the *no ruler* condition children's estimates appeared to reflect prior feedback during training – which was not likely to transfer across spatially-transformed number lines or be maintained over time. Given their lack of correct estimation of the midpoint, when verbally probed, it is unlikely that these children spontaneously discovered and applied a midpoint-based landmark strategy. Thus, while these children did not revert to a logarithmic representation on subtests, their accuracy – in terms of linearity and means PAE – decreased substantially from the training task over the four post-subtests.

On the other hand, children in the *incongruent ruler* condition, by utilizing information depicted on the ruler, were less likely to focus on non-transferable cues. For this reason they were equally likely to have correct estimates on the first trial of each training block – where a large delay likely cleared working memory of the previous trial's feedback – as the subsequent seven trials in a block. Furthermore, the generality of the midpoint-landmark strategy provided the children with a means to generate a spatial reference across all subtests.

Analysis of the magnitude over time curves provided tentative evidence that, in comparison to children in the *no ruler* condition, the children in the *incongruent ruler* and the *congruent ruler* conditions were more likely to stop at the midpoint; however the lack of distinction between the *congruent ruler* and the *incongruent ruler* conditions requires additional study, perhaps with more incentive to use the mouse to express strategy or by bypassing the mouse all together. Specifically, in some cases children

referenced the landmark visually and stated (approximately), “It is right there past 90”. Therefore, eye-tracking following training would be likely to reveal attention to standard reference points (e.g. Sullivan, Juhasz, Slattery, & Barth, 2011). To a lesser extent, a touch-sensitive computer screen might prompt more explicit use of landmark references – particularly if the hand occluded a clear view of the number line.

Temporally dynamic analysis, such as mouse-tracking, provided an implicit quantitative window into the strategy of an individual as he or she engaged the task. While a child could be questioned about his or her strategy, there is no assurance that the strategy was applied consciously, could be accurately recalled, or would not be invented post-hoc to meet demand characteristics. Furthermore, if the child was questioned during the task, then the child’s behavior could be affected on future trials. Yet, some anecdotal observations suggest that children were consciously aware of having applied a strategy. For example, a student from the *incongruent ruler* condition stated, “I just imagine where forty-five, ninety, and one hundred thirty-five is.” Likewise, another child stated that, “Forty-five is between zero and ninety, and ninety is in the middle” as she navigated the cursor according to these landmarks. A strong study of children’s estimation strategies would compare both verbal accounts and implicit, quantitative methods.

While the results suggest that children often did use an explicit strategy of seeking the midpoint, fewer children did the same for the first quartile landmark (45), and even fewer for the third quartile landmark (135). Specifically, while the trial curve analysis revealed some results of peaks (stops) near 90, a preliminary analysis for peaks near 45 and 135 did not reveal differences between conditions [i.e., distributions of stops at 45 for targets 36 and 56 did not differ between conditions significantly, $X^2(2) = 2.7$, $p > .2$; likewise, distribution of stops at 135 for targets 120 and 155 did not differ between conditions significantly , $X^2(2) = 1.5$, $p > .2$].

Considering this lack of evidence for an explicit strategy for 45 and 135, as well as the difficulty children encountered in the corresponding bisection questions in the second experiment, it seems that the children conceptualized the midpoint differently than the quarter points. While, the intention of instruction was to promote both the spatial and numerical significance of the quarter points as the halves of halves, children rarely displayed an understanding that 45 and 135 were chosen because of their

numerical significance. Rather, children may have viewed the values 45 and 135 as numerically-arbitrary. As discussed previously, this may have arisen from these particular numbers oddness.

The clearest example of this inability to make use of the numerical significance of these quartile points comes from the numerical transfer subtest, in which a number of children incorrectly estimated 45 near the quarter-mark of the 0 – 90 number line. Rather than learning to associate 45 with one-quarter of the 0 – 180 scale, they appeared to associate 45 with the quarter mark of the spatial display, only. When facing the new 0 – 90 scale these children retained the spatial significance but not the numerical significance. As explained in the discussion of experiment 2, this often led to a pattern of curves that was best-fit by an exponential function. To some extent this result was likely dependent on the younger age of the children in experiment 2 (early 2nd graders). Experiment 1, with slightly older children, did not include a numerical-transfer task. However, the misconception likely arose from the strategies that children applied while using the ruler. Possible remediations are discussed below.

5.2. Strategy vs. representation

While the previous section discussed the application of strategy grounded on the use of specific landmarks, previous research on number line estimation typically describes estimation performance in terms of the quality of the underlying representation (Barth & Paladino, 2011; Izard & Dehaene, 2008; Siegler & Opfer, 2003). According to Dehaene (1997), all estimation tasks are grounded by a mental number line that encodes a general, amodal mapping between space and number. As Siegler and Opfer (2003) describe, the logarithmic encoding of this number line is manifested in the estimates of young children, while adults tend to shift towards a more normative, linear representation. Siegler and colleagues assert that the quality of this underlying representation influences performance on a number of critical mathematical tasks (Booth & Siegler, 2006; Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009).

At least superficially, development of a representation and development of a strategy are quite different. These concepts reflect the traditional distinction between declarative or conceptual knowledge (propositions) and procedural knowledge (skills), which have been described as cognitively and neurologically distinct (Anderson, 1983). Unlike declarative knowledge, which can be applied quit

flexibly, procedural knowledge is often situated to specific tasks and goals (Anderson, 1993). From this perspective, is there a benefit to teaching an explicit strategy for estimation that might only be relevant to the given task? Indeed, numerical and spatial bisection is a critical strategy in a number of tasks, such as fraction estimation and arithmetic (Siegler, Thompson, & Schneider, 2011) and geometry (Vitale et al., 2010), and therefore deserves at least some curricular attention.

Yet, the significance of the number line as an instructional tool lies in its potential to affect the underlying sense of numerical magnitude, which influences performance across nearly all mathematical tasks. Does learning a heuristic strategy allow the learner to bypass development of the underlying numerical magnitude representation? Perhaps, but the alternative of providing no strategy or direction is untenable. Children in the *no ruler* condition, who were not taught any specific strategy, discovered their own context-sensitive cues that served well-enough to reach criterion, but not for transfer.

While there may be a drawback to focusing on the instruction of procedural, or strategic skills exclusively, the relationship between strategy and representation is often reciprocal (Star, 2012). In a study of where children manipulated decimal fractions on a number line Rittle-Johnson, Siegler, and Alibali (2001) found an iterative relationship between procedural and conceptual knowledge, such that providing strategies to students would then improve their concept of decimal fraction magnitude, and vice-versa. In the case of this study, while the numerical transfer subtest exceeded the children's current capabilities as an assessment, given feedback the same task would have been highly beneficial as an additional training task. A clear drawback to the instructional approach was the lack of variety in the training. Future studies should look to incorporate greater diversity in the spatial and numerical qualities of the training display. By applying a well-rehearsed landmark-based strategy to a number of different number lines, the underlying representation of magnitude is likely to develop in parallel.

Yet, this defense of strategic instruction rather than conceptual representation-based instruction, begs the question of whether these two forms of knowledge are meaningfully distinct. Can mathematical representations exist without procedures and strategies for their use? While the underlying, implicit mental number line may exist in a relatively static form (Dehaene, 1997; Piazza et al., 2004), the process of applying it to numerical tasks likely requires some amount of deliberative strategy. For example, Izard and Dehaene (2008) propose a model of numerical estimation in which the mental number line is

encoded in a logarithmic format, but requires an affine transformation (scale and shift) to be used. This transformation process is likely to be at least partially explicit and deliberative. According to this model, except in circumstances where an individual is under tremendous time pressure or is untrained, application of the mental number line to meet the needs of linear tasks involves deliberative processing.

Therefore, is the "linear representation" described by Siegler and Opfer (2003) a static representation like the mental number line described by Dehaene (1997) if it requires explicit calibration to execute? Or rather, is the linear representation a higher-level concept that integrates both implicit and explicit forms of knowledge? From this perspective, the strategies developed here to assist children in estimating may be an integral part of the representation, itself. This idea is conveyed by Barth and Paladino's (2011) model where maturation is accomplished by subdividing the numerical scale at its midpoint (and potentially quartiles). In particular, this model does not address any static, implicit mental number line, but portrays estimation in a more dynamic sense. Likewise, even with the traditional linear model, as applied to fraction estimation, deliberative action takes a central role, and is described as a component of the child's representation (Siegler et al., 2011).

While this study does not offer conclusive evidence for either the log-linear or power models (the distributions of best-fitting models were similar), the data here does suggest that individual's strategy can have a significant influence on patterns of estimates. The intimate link between strategy and representation justifies instruction focused on the development a particular strategy. However, if a specific strategy is to be instructed, then it should be spatially- and numerically-meaningful, intuitively comprehensible, and highly flexible. A quartile landmark-based strategy meets these criteria. Future research is needed to answer questions about the necessity and sufficiency of this strategy and how it can be applied to a wider range of estimation and calculation tasks.

5.2. Instructional design implications

Concrete, spatial tools offer learners an opportunity to actively participate with the underlying mechanisms of a concept (Mix, 2009). This interaction can facilitate a conceptualization that is more deeply interwoven with the learner's prior concepts and intuitive processes. More importantly, a well-

designed tool can foster the adaptation of prior concepts to fit new experiences. Designing for deep change in children's concepts, is a central challenge for instructional tools.

With the emergence of digital tools a number of instructional options are available, including: real-time feedback and assessment, visual depictions of concepts, and gestural interaction with virtual objects. While it is appealing to implement as many assistive elements as possible in the design of an instructional tool, there can be "too much of a good thing". For example, research on cognitive tutors finds that children often take advantage of feedback systems to navigate through lessons without truly interacting with the concepts (Baker, Corbett, & Koedinger, 2004). On the other hand, too little guidance may overwhelm the learner (Kirschner et al., 2006) or foster misconceptions (Simmons & Cope, 1990).

This study confirms that both ends of the instructional spectrum limit learning. In terms of minimal instruction the *no ruler* condition was designed to facilitate independent discovery of the given numerical representation, and applicable strategies. As predicted, these children, on average, required significantly more trials to achieve the same criterion as the other conditions. Yet, this added cost of learning efficiency would be well worth it if they had developed highly robust representations of the numerical concept – particularly considering that these children required less initial instructional time. However, the posttest task conveyed the context-specificity and time-sensitivity of their knowledge, given the decayed performance as the task was spatially modified. For example, many of these children relied upon the feedback of the previous trial to serve as a reference for the following trial.

Yet, one might argue that this finding reflects a specific artifact of the task's design. If the software had been designed in another way – for example, by displaying feedback only after a block of trials – this non-adaptive strategy would not have emerged. Therefore, while this condition was claimed to be minimal guidance, the source of its problem was too much guidance. While this is certainly a possibility, by reducing the amount of feedback any further I risked the possibility that a number of children would not have reached criterion at all. This level of inefficiency is unrealistic in contemporary educational settings.

Therefore, rather than attempting to reduce feedback in the face of (unintended) misuse, it is more appropriate to use instructional devices to direct students attention and behavior. By providing a tool that transparently and directly displays the target concepts and strategies, with little room for error,

we can guard against the development of unexpected misconceptions. This was the goal in the design of the *congruent ruler*. This tool displays a representation of the number line that is context-general, conveys a specific strategy (divide-and-conquer), and easy to manipulate. However, as seen in the results of both experiments, the intuitive design of the tool, which fosters immediately strong performance, comes at the cost of retention and transfer.

This result fits well within the framework of desirable difficulties (E. L Bjork & R. A. Bjork, 2011; R. A. Bjork, 1994; R. A. Bjork & Linn, 2006). In memory studies, the benefit of desirable difficulties often lies in the difference between forgetting and re-encoding a memory, which strengthens its associations with other memories, and applying an activate memory, which has little affect on its encoding (Bjork, 1994; Christina & Bjork, 1991). For example, if a learning sequence is spaced across long delays, then the learner is likely to forget and re-encode information, encoding new features of the learning context. On the other hand, if learning is massed in a single session, then the memory representation is more likely to reflect context-specific aspects of the single encounter with the concept.

In the context of instructional technology, a constant stream of feedback and assistance relinquishes the learner of his or her responsibility to engage the task independently – interpreting meaning of elements, recalling relevant information, modifying strategy as a consequence of feedback, etc. In the case of the *congruent ruler* the persistent, one-to-one correspondence between ruler and target number line – which most children maintained by either holding the ruler up to the on-screen number line or placing it directly below – did nothing to prompt questions about the meaning of the landmarks. Specifically, because children were focused on the objective of the game (catching the fish in a minimum of trials), reflection on the numerical and spatial relationships between the chosen landmarks and the numerical scale as a whole was superfluous. However, active and explicit reflecting upon relationship between conceptual components is necessary to facilitate the construction of a robust, integrated representation of the concept (Linn, 2000)

In terms of instruction, in general, this result implies that robust learning occurs only when the learner is willing to take on conceptual challenges. This is often associated with performance deficits and a feeling of having learned little (Kornell & Bjork, 2009). Inversely, consistently high performance during a learning task should be looked upon with caution. For example, while an instructor may design a lesson

to be dynamic, accessible, culturally-relevant, etc., thereby facilitating high participation and strong in-task performance, he or she may be surprised to discover poor outcomes once the high-interest elements of the learning context have been removed. On the other hand, if the instructor can assist in helping his or her students make the content meaningful to themselves – a much more demanding process – they are more likely to retain the information.

In both experiments described here, this position was implemented through the *incongruent ruler*. Rather than giving participants a tool whose usage was fully implied, the *incongruent ruler* required an interpretation. Children were told that the ruler could be used effectively; however, they were responsible for doing so. Beyond initial interpretation, whatever process the children applied to project landmarks from the ruler to the number line was likely re-engaged frequently. This contrasts with the *congruent ruler*, where the position was set during the initial trial and rarely changed. This difference between materials, and the behaviors they afforded, corresponds well to the idea of the value of forgetting.

Another advantage of the *incongruent* ruler*,* which differentiated it from the *no ruler condition*, was that while some independent interpretation of the task was necessary, the breadth of this process was limited. There were only a handful of ways that children could have effectively applied the tool. While individual behavior, outside the context of the software, was not recorded (beyond some noted anecdotes) the experimenter observed two general patterns in participant strategies. In some cases children attempted to locate the target magnitude on the ruler directly, and then translate this location to the on-screen number line. For example, with a target of 60, the child would place his or her finger at a location between 45 and 90 on the ruler, and then attempt to visualize this location on the on-screen number line. In other cases the children used the ruler to locate landmarks on the number line, and then used these (imagined) landmarks to guide final estimates. For example, with a target of 60, the child would use the ruler as a reference for the 45 and/or 90 landmarks, locate these references on the number line, and then proceed to estimate a magnitude between these references.

While these two strategies appear equally valid for achieving criterion in the learning task, the former – direct projection from ruler to number line – was less likely to promote transfer at posttest. Like the *congruent* ruler this approach bypassed the need to internalize the numerical and spatial significance of the chosen landmarks. Once these children reached the posttest, without access to the ruler, they no

longer had a reference for these landmarks. On the other hand, by actively locating the landmarks on the target number line, children eventually internalized their location, and were able to apply this well-developed representation during posttest. In this study there is no obvious way to tease apart these two (and possible other) strategies. Future work may address the specific strategies that children apply by including additional constraints in the use of materials. For example, if the ruler was only available during the first second of a trial, and then removed, the child would be more likely to apply a more flexible strategy.

Despite the potential for inappropriate strategies, the *incongruent ruler* was an effective implementation of the main theoretical position of this manuscript: that limited challenges to the coordination of materials embedded in a learning task can foster more robust concepts. However, to some degree the particular tool chosen was arbitrary. Other versions of the *incongruent ruler* were piloted with children, but dismissed because of the propensity to develop non-adaptive strategies. For example, a shortened version of the ruler was constructed as the original incongruent ruler; however, a number of children held this shortened version of the ruler a fixed distance from the screen so that its projected image fit the on-screen number line congruently. In a sense these children had constructed the *congruent ruler* spontaneously. Other variations of this tool that were considered included: a ruler whose orientation was fixed vertically, a ruler whose position was fixed at a large distance from the on-screen number line, a digital ruler that could not be viewed simultaneously with the target number line, a digital ruler that visually faded from view after a number of correct trials, and a ruler with an alternate shape (e.g. a circular ruler).

I suspect that each of these manipulations, if they did not foster unintended behaviors, would have afforded similar gains in learning. However, it may also be the case that the benefit of the *incongruent ruler* (as applied here) stemmed purely from the difference in scale from the on-screen number line. Specifically, given that two of the four subtests utilized number lines with a transformed scale, engaging in a similar process during training may have prepared children to perform this task at the posttest. However, there was no indication that *incongruent ruler* participants had any more advantage on subtest 2 (half-scale) then subtest 3 (full-scale, reversed).

Clearly, designing materials that do not afford the use of non-adaptive strategies is a challenge even if one is committed to limiting the amount of instruction and feedback. Children are highly adept at finding context-specific strategies to solve problems in a learning environment – particularly a computer-based learning environment with feedback (Miller, Lehman, & Koedinger, 1999). Therefore studies of learning with digital tools should ensure that comprehensive records of strategy-relevant behaviors are taken. While this was attempted here through mouse-tracking, it was clear that much of the child's strategy occurred offline.

Another limitation of the incongruent ruler was that children appeared to make very little use of the 45 or 135 landmarks at posttest. Poor results in the associated bisection questions, across all conditions, suggest that the children simply did not interpret their numerical significance – even though their meaning was explained during instruction (i.e., "halfway between 0 and 90"). While perhaps not critical for this task, being able to understand the significance of a quarter (or any division of a quantity) is a critical part of fraction comprehension. Simply presenting the children with a visual depiction of a quartered number line, along with a short verbal explanation, was clearly insufficient for promoting a strong understanding. Alternatively, given the importance of grounding abstract concepts in physical behaviors (Black et al., 2012), a child might be asked to generate landmarks by folding (or cutting) a number line in half, and then folding each half in half to generate quartiles.

In conclusion, the manipulation in this experiment clearly supports the hypothesis that learning tools must strike a balance between overly limited and overly limiting guidance and feedback. While independent behavior and thinking is critical to learning, without constraint the learner is likely to pick up non-adaptive strategies. Planning for unintended consequences is a tremendous challenge for instructional design, and requires deep understanding of the conceptual domain and repeated testing with children. While the tool applied here was effective, flaws were inevitable. Yet, desirable difficulties are equally as relevant to the professional researcher as they are for the primary school student.

REFERENCE

Akaike, H. (1974). A new look a the statistical model identification. *IEE Transactions on Automatic Control*, *19*(6), 716-723.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics 30*, *A*, 9-14.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., Reder, L. M., Simon, H. A., Ericsson, K. A., & Glaser, R. (1998). Radical Constructivism and Cognitive Psychology. *Education*, *1*(1), 227-278.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *7*(3), 522-30.

Anderson, M. C., Neely, J. H., Bjork, E. L., & Bjork, R. (1996). Interference and inhibition in memory retrieval. *Memory* (pp. 237-313).

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems, *Intelligent tutoring systems* (Vol. 4053, pp. 531-540).

Ball, D. L. (1992). Magical hopes: Manipulatives and the reform of math education. *American Educator*, *16*(1), 14-19.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *358*(1435), 1177-1187.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617-645.

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science*, *14*(1), 125-135.

Barth, H. C., Slusser, E., Cohen, D., & Paladino, A. M. (2011). A sense of proportion: commentary on Opfer, Siegler and Young. *Developmental Science*, *14*(5), 1205-1206.

Bjork, Elizabeth L, & Bjork, R. A. (2011). Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 55-64). New York: Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A., & Linn, M. C. (1999). Introducing Desirable Difficulties for Educational Applications in Science (IDDEAS). *Environment*, 1-28.

Bjork, R. A., & Linn, M. C. (2006). The science of learning and the learning of science: Introducing desirable difficulties. *APS Observer*, *19*(3), 29-39.

Black, J. B., Segal, A., Vitale, J. M., & Fadjo, C. L. (2012). Embodied cognition. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments* (2nd ed., pp. 198-223). New York: Routledge.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: the roles of structural and superficial similarity. *Memory & cognition*, *28*(1), 108-24.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189-201.

Brainerd, C. J. (1972). Reinforcement and reversibility in quantity conservation acquisition. *Psychonomic Science*, *27*, 114-116.

Brown, M. C., McNeil, N. M., & Glenberg, A. M. (2009). Using concreteness in education: Real problems, potential solutions. *Child Development Perspectives*, *3*(3), 160-164.

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Case, R., & Griffin, S. (1990). Child cognitive development: The role of central conceptual structures in the development of scientific and social thought. In C. A. Hauret (Ed.), *Developmental psychology: Cognitive, perceptuo-motor, and neurophysiological perspectives* (pp. 193-230). North-Holland: Elsevier.

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., Stephenson, K. M., et al. (1996). The Role of Central Conceptual Structures in the Development of Children's Thought. *Monographs of the Society for Research in Child Development*, *61*(1/2), i.

Chariker, J. H., Naaz, F., & Pani, J. R. (2011). Computer-based learning of neuroanatomy: A longitudinal study of learning, transfer, and retention. *Journal of Educational Psychology*, *103*(1), 19-31.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152.

Christina, R. W., & Bjork, R. A. (1991). Optimizing long-term retention and transfer. In D. Druckman & R. A. Bjork (Eds.), *In the mind's eye: Enhancing human performance* (pp. 23-56). Washington, DC: National Academy Press.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345-351.

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *Journal of the Learning Sciences*, *12*(4), 451-493.

Dehaene, S, & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, *1*(1), 83-120.

Dehaene, S, & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1-29.

Dehaene, Stanislas. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Dehaene, Stanislas, Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3/4/5/6), 487-506.

Dehaene, Stanislas, Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371-396.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., et al. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428-1446.

Duncker, K. (1945). On problem-solving. *Psychologcial Monographs*, *58*(Whole No. 270).

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year old children: Evidence for a segmented linear model. *Journal of Experimental Child Psychology*, *99*(1), 1-17.

Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007). Tracking the continuity of language comprehension: computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, *31*(5), 889-909.

Fauconnier, G, & Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, *22*(2), 133-187..

Fauconnier, Gilles. (1994). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. *Language* (Vol. 63, p. 142).

Fauconnier, Gilles, & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. *EUA Basic Books* (p. 440).

Fay, A. L., & Mayer, R. E. (1994). Benefits of teaching design skills before teaching LOGO computer programming: Evidence for syntax independent learning. *Journal of Educational Computing Research*, *11*, 187-210.

Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Science*, *8*(7), 307-314.

Fennema, E. H. (1972). Models and Mathematics. *Arithmetic Teacher*, *20*(4), 635-640.

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitudes. *Journal of Cognitive Neuroscience*, *15*(1), 47-56.

Fuson, K. C., & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first- and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *21*(3), 180-206.

Gagné, R. M., & Brown, L. T. (1961). Some factors in the programming of conceptual learning. *Journal of experimental psychology*, *62*(4), 313-321.

Gelman, R. (1969). Conservation Acquisition : to Attend to Relevant Problem of Learning. *Journal of Experimental Child Psychology*, *7*, 167-187.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. (D. Gentner & A. L. Stevens, Eds.)*Cognitive Science*, *7*(2), 155-170.

Gentner, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752-775.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analagocial transfer. *Cognitive Psychology*, *15*, 1-38.

Glenberg, A. M. (1999). Perceptual symbols in language comprehension. *Behavioral and Brain Sciences*, *22*(4), 618–619.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*(2-3), 231-262.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the learning sciences*, *14*(1), 69-110.

Goldstone, R. L., & Son, J. Y. (2008). A well grounded education: The role of perception in science and mathematics. *Symbols embodiment and meaning*, 327–355.

Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *The Journal of the Learning Sciences*, *17*(4), 465-516.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, *2*(2), 265-284.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: guiding attention guides thought. *Psychological Science*, *14*(5), 462-466.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 25-49). Cambridge, MA: MIT Press.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with math achievement. *Nature*, *455*(7213), 665-668.

Hays, M. J., Kornell, N., & Bjork, R. a. (2010). The costs and benefits of providing feedback during learning. *Psychonomic bulletin & review*, *17*(6), 797-801.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Human, P., Murray, H., Olivier, A., et al. (1996). Problem Solving as a Basis for Reform in Curriculum and Instruction: The Case of Mathematics. *Educational Researcher*, *25*(4), 12-21.

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, *107*(3), 500-524.

Hollands, J. G., Tanaka, T., & Dyre, B. P. (2002). Understanding bias in proportion production. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(3), 563-574.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: the numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology*, *103*(1), 17-29.

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, *6*(6), 435-448.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221-1247.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science*, *25* (April), 454-455.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimally guided instruction does not work. *Educational Psychologist*, *41*, 75-86.

Kittel, J. E. (1957). An experimental study of the effect of external direction during learning on transfer and retention of principles. *Journal of Educational Psychology*, *48*, 391-405.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of experimental psychology: General*, *138*(4), 449-468.

Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. *Basic Books* (p. xvii, 493). New York: Basic Books.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life.* New York: Cambridge University Press.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Lee, M., & Thompson, A. (1997). Guided instruction in LOGO programming and the development of cognitive monitoring strategies among college students. *Journal of Educational Computing Research*, *16*, 125-144.

Linn, M. C. (2000). Designing the knowledge integration environment. *International Journal of Science Education*, *22*(8), 781–796.

Linn, M. C., Chiu, J., Zhang, H., & McElhaney, K. (2010). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. Benjamin (Ed.), *Successful remembering and successful forgetting a Festchrift in honor of Robert A Bjork* (pp. 1-35).

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science*, *14*, 396-401.

Loman, N. L., & Mayer, R. E. (1983). Signaling techniques that increase the understandability of expository prose. *Journal of Educational Psychology*, *75*(3), 402-412.

Lorch, R. F. J., & Lorch, E. P. (1995). Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology*, *87*(537-544), 537-544.

Lorch, R. F., Lorch, E. P., & Inman, W. E. (1993). Effects of signaling topic structure on text recall. *Journal of Educational Psychology*, *85*(2), 281-290.

Lorch, R. G., & Lorch, E. P. (1996). Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology*, *88*(1), 38-48.

Lourenco, S. F., & Longo, M. R. (2009). Multiple spatial representations of number: evidence for co-existing compressive and linear scales. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, *193*(1), 151-6.

Mannes, S. M., & Kintsch, W. (1987). Knowledge Organization and Text Organization. *Cognition and Instruction*, *4*(2), 91-115.

Martin, T., & Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science*, *29*(4), 587-625.

Martin, T., Lukong, A., & Reaves, R. (2007). The role of manipulatives in arithmetic and geometry. *Journal of Educational and Human Development*, *1*(1).

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a Cognitive Guide in Multimedia Learning. *Journal of Educational Psychology*, *93*(2), 377-389.

Mautone, P. D., & Mayer, R. E. (2007). Cognitive aids for guiding graph comprehension. *Journal of Educational Psychology*, *99*(3), 640-652.

Mayer, R. E. (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? *American Psychologist*, *59*(1), 14-19.

McNeil, N. M., & Jarvin, L. (2007). When theories don't add up: Disentangling the manipulatives debate. *Theory Into Practice*, *46*(4), 309-316.

McNeil, N. M., & Uttal, D. H. (2009). Rethinking the Use of Concrete Materials in Learning: Perspectives From Development and Education. *Child Development Perspectives*, *3*(3), 137-139.

Meira, L. (1998). Making sense of instructional devices: The emergence of transparency in mathematical activity. *Journal for Research in Mathematics Education*, *29*(2), 121-142.

Meyer, B. J. F. (1975). *The organization of prose and its effect on memory*. Amsterdam: North Holland.

Miller, C., Lehman, J., & Koedinger, K. R. (1999). Goals and learning in microworlds. *Cognitive Science*, *23*(3), 305-336.

Mix, K. S. (2009). Spatial tools for mathematical thought. *The Spatial Foundations of Language and Cognition*, *1*(9), 40–66.

Moyer, P. S. (2001). Are we having fun yet? How teachers use manipulatives to teach mathematics. *Educational Studies in Mathematics*, *47*(2), 175-197.

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist*, *44*(1), 20-40.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169-195.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*(3), 788-804.

Opfer, J. E., Siegler, R. S., & Young, C. J. (2011). The powers of noise-fitting: reply to Barth and Paladino. *Developmental Science*, *14*(5), 1194-1204.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. *New York* (p. 246). Basic Books.

Pea, R. D., & Kurland, D. (1984). On the cognitive effects of learning computer programming. *New Ideas in Psychology*, *2*(2), 137-168. Ablex Publishing Corp.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Piaget, J. (1954). *The construction of reality in the child*. (T. Béla, K. Janó, & Z. Afasz, Eds.)*The construction of reality in the child* (p. 386). Basic Books.

Piaget, J. (1962). *Play, dreams, and imitation in childhood*. New York: Norton.

Piaget, J. (1970). *Science of education and the psychology of the child*. New York: Orion Press.

Piazza, M., Izard, V. V., Pinel, P., Bihan, D. L., Dehaene, S., & Le Bihan, D. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547-555.

Pinel, P., Piazza, M., Bihan, D. L., & Dehaene, S. (2004). Distributed and overlapping cerebral represenations of number, size, and luminance during comparative judgments. *Neuron*, *41*, 1-20.

Quinn, P. C., & Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology*, *1*(3), 271-287.

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, *79*(2), 375-394.

Richland, L., Linn, M. C., & Bjork, R. A. (2007). Cognition and Instruction: Bridging Laboratory and Classroom Settings. In F. Durso (Ed.), *Handbook of Applied Cognition Second Edition* (pp. 555-583).

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346-362.

Sarama, J., & Clements, D. H. (2009). "Concrete" computer manipulatives in mathematics education. *Child Development Perspectives*, *3*(3), 145-150.

Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., & Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. *Journal of Educational Psychology*, *102*(4), 872-879.

Schwartz, D. L., Varma, S., & Martin, T. (2008). Dynamic transfer and innovation. In S. Vosniadou (Ed.), *Handbook of Conceptual Change* (Vol. 95616, pp. 479-506). Mahwah, NJ: Routledge.

Shulman, L. S., & Keisler, R. E. (1966). *Learning by discovery*. Chicago: Rand McNally.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428-444.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237-243.

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental science*, *11*(5), 655-61.

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, *101*(3), 545-560.

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The Logarithmic-To-Linear Shift : One Learning Sequence , Many Tasks , Many Time Scales. *Society*, *3*(3), 143-150.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive psychology*, *62*(4), 273-96.

Simmons, M., & Cope, P. (1990). Fragile knowledge of angle in Turtle Geometry. *Educational Studies in Mathematics*, *21*, 375-382.

Son, J. Y., & Goldstone, R. L. (2009a). Contextualization in perspective. *Cognition and Instruction*, *57*(1), 51-89.

Son, J. Y., & Goldstone, R. L. (2009b). Fostering general transfer with specific simulations. *Pragmatics Cognition*, *17*(1), 1-42.

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207-211.

Star, J. R. (2012). Procedural Knowledge Reconceptualizing. *Journal for Research in Mathematics Education*, *36*(5), 404-411.

Storm, B. C., Angello, G., & Bjork, E. L. (2011). Thinking can cause forgetting: Memory dynamics in creative problem solving. *Journal of Experimental Psychology: Learning, memory and cognition*, *37*(5), 1287-1293.

Sullivan, J. L., Juhasz, B. J., Slattery, T. J., & Barth, H. C. (2011). Adults' number-line estimation strategies: Evidence from eye movements. *Psychonomic bulletin & review*, *18*(3), 557-63.

Thelen, E. (2000). Grounded in the World: Developmental Origins of the Embodied Mind. *Infancy*, *1*(1), 3-28.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768-1786.

Uttal, D. H., Doherty, K. O., Newland, R., Hand, L. L., & Deloache, J. (2009). Dual Representation and the Linking of Concrete and Symbolic Representations. *Child Development*, *3*(3), 156-159.

Viarouge, A., Hubbard, E. M., Dehaene, S., & Sackur, J. (2010). Number line compression and the illusory perception of random numbers. *Experimental Psychology*, *57*(6), 446-454.

Vitale, J. M., Black, J. B., Carson, E., & Chang, C. (2010). Development in the Estimation of Degree Measure: Integrating Analog and Discrete Representations. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2242-2247). Austin, TX: Cognitive Science Society.

Vitale, J. M., Siegler, R. S., & Black, J. B. (2011). Promoting the development of linear numerical representations through a video game-based intervention. Montreal, QC, Canada,.

Wallach, L., & Sprott, R. L. (1964). Inducing number conservation in children. *Child Development*, *35*, 71-84.

Xu, F., & Spelke, E. (2000). Large number discrimination in 6-month old infants. *Cognition*, *74*, B1-B11.

Appendix A – Training Task ANOVAs and ANCOVAs

ANOVA – Total blocks (Training)

Source	SS	df	MS	F	η_p^2
Condition	304	2	152	13.5 ***	.26
Congruent ruler vs. others	219	1	219	19.4 ***	.20
No ruler vs. Incongruent	88	1	88	7.8 **	.09
Error	869	77	11		

*** p < .001 ** p < .01 * p < .05 † p < .1

ANCOVA – Total blocks (Training)

Source	SS	df	MS	F	η_p^2
Condition	319	2	152	14.9 ***	.28
Congruent ruler vs. others	214	1	214	20.0 ***	.21
No ruler vs. Incongruent	102	1	102	9.5 **	.11
Age (in months)	57	1	57	5.3 *	.07
Error	812	76	10.7		

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Total duration (Training)

Source	SS	df	MS	F	η_p^2
Condition	662000	2	331000	3.3 *	.08
Congruent ruler vs. others	326000	1	326000	3.3 †	.04
No ruler vs. Incongruent	329000	1	329000	3.3 †	.04
Error	773000	77	100000		

*** p < .001 ** p < .01 * p < .05 † p < .1

(post hoc comparison) *Congruent ruler vs. incongruent ruler*: p = .05.

ANCOVA – Total duration (Training)

Source	SS	df	MS	F	η_p^2
Condition	512000	2	256000	3.0 †	.07
Age (in months)	1190000	1	1190000	13.9 ***	.15
Error	6540000	76	86000		

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Total duration (Training)

Source	SS	df	MS	F	η_p^2
Condition	662000	2	331000	3.3 *	.08
Congruent ruler vs. others	326000	1	326000	3.3 †	.04
No ruler vs. Incongruent	329000	1	329000	3.3 †	.04
Error	773000	77	100000		

*** p < .001 ** p < .01 * p < .05 † p < .1

(post hoc comparison) *Congruent ruler vs. incongruent ruler*: p = .05.

ANOVA – Overall Mean PAE (Training)

Source	SS	df	MS	F	η_p^2
Condition	.008	2	.004	7.4 **	.16
Congruent ruler vs. others	.005	1	.005	9.6 **	.11
No ruler vs. Incongruent	.003	1	.003	5.0 *	.07
Error	.041	77	.001		

*** p < .001 ** p < .01 * p < .05 † p < .1

(post hoc comparison) *Congruent ruler vs. incongruent ruler*: p = .05.

ANOVA – Overall Percent Trials Correct (Training)

Source	SS	df	MS	F	η_p^2
Condition	12.2	2	6.1	8.7 ***	.19
Congruent ruler vs. others	7.4	1	7.4	10.5 **	.04
No ruler vs. Incongruent	4.9	1	4.9	6.8 *	.04
Error	53.6	77	.70		

*** p < .001 ** p < .01 * p < .05 † p < .1

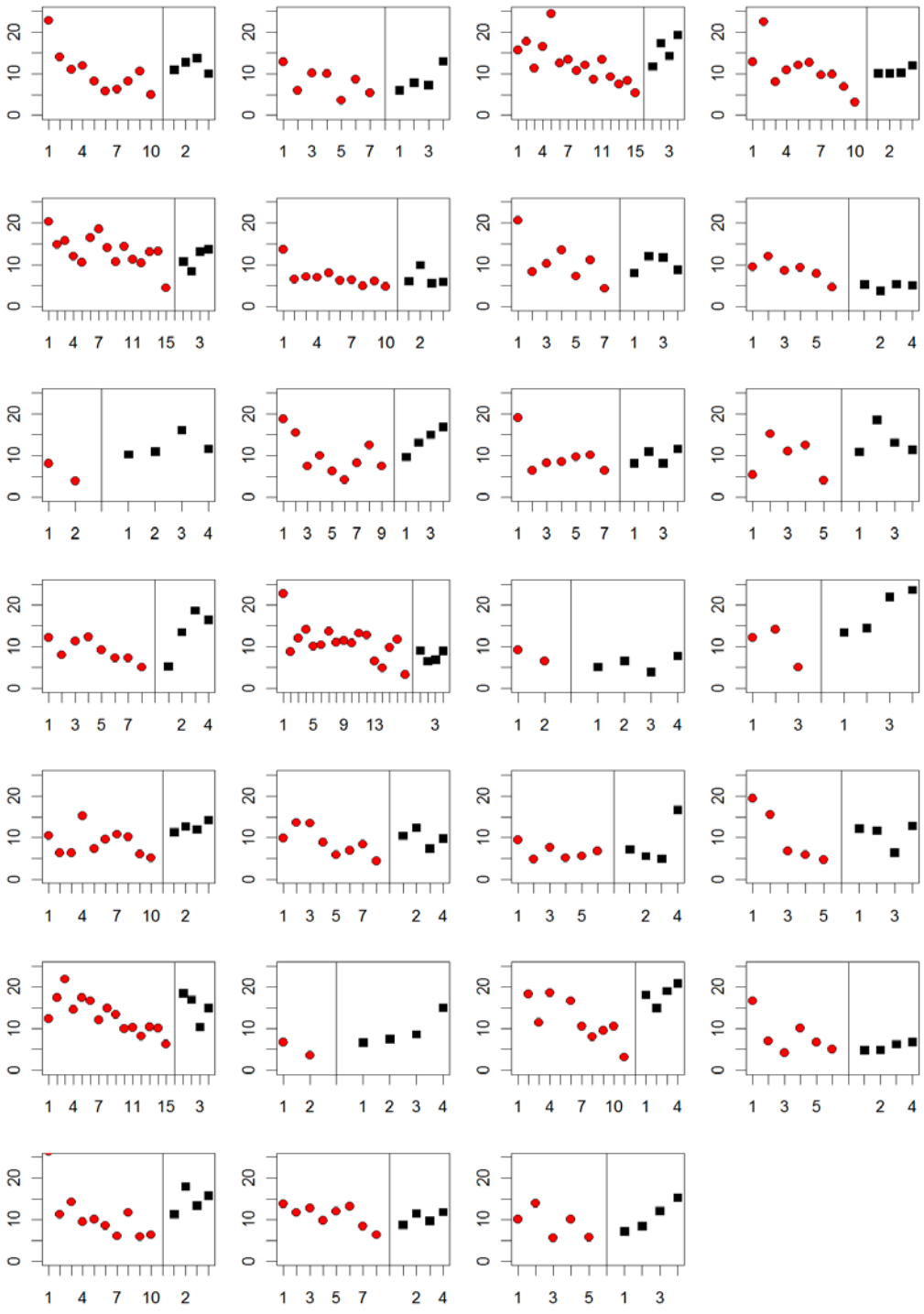
ANOVA – Block 1 Percent Trials Correct (Training)

Source	SS	df	MS	F	η_p^2
Condition	24.5	2	12.3	5.1 **	.12
Congruent ruler vs. others	15.2	1	15.2	6.3 **	.08
No ruler vs. Incongruent	9.2	1	9.2	3.9 *	.05
Error	183.5	77	.70		

*** p < .001 ** p < .01 * p < .05 † p < .1

Appendix B – 1

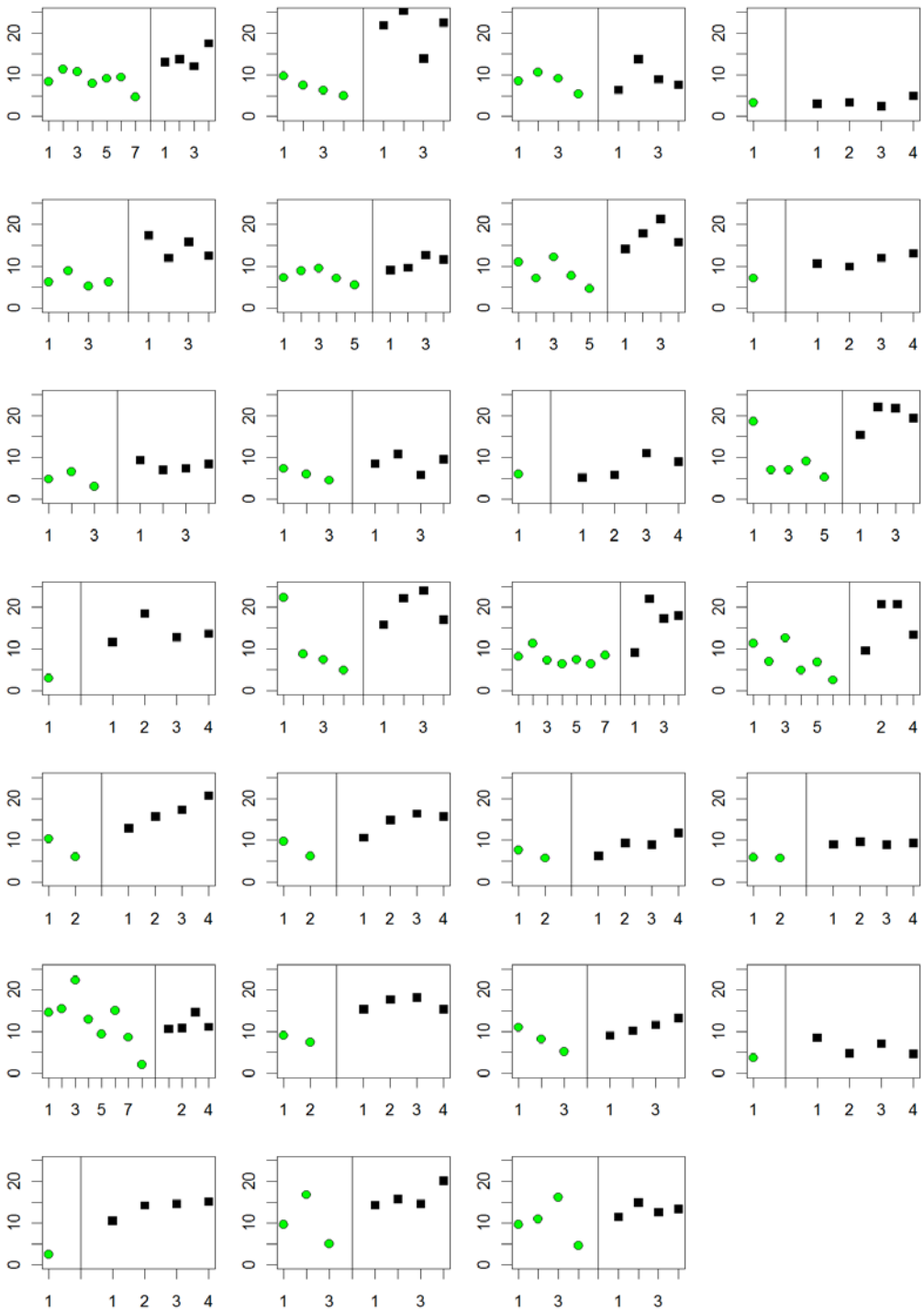
Mean PAE across all Training Blocks and Post-subtests – *No ruler* Condition



Plots of individual participants in the *no ruler* condition show progression of mean errors through all training blocks (consisting of 8 trials each) and subtests (consisting of 19 trials each). Red points indicate training blocks, while black squares indicate post-subtests.

Appendix B-2

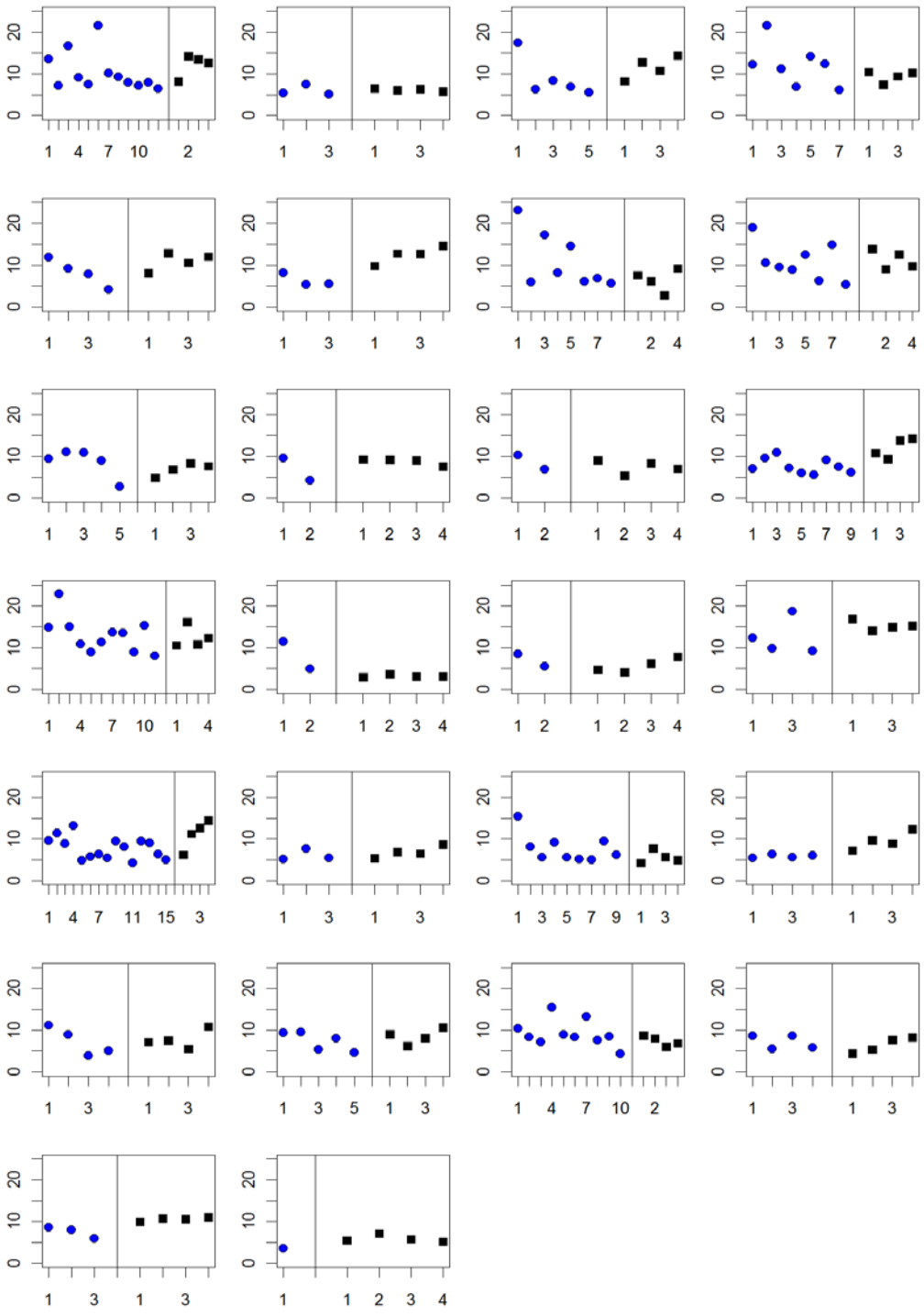
Mean PAE across all Training Blocks and Post-subtests – *Congruent ruler* Condition



Plots of individual participants in the *congruent ruler* condition show progression of mean errors through all training blocks (consisting of 8 trials each) and subtests (consisting of 19 trials each). Green points indicate training blocks, while black squares indicate post-subtests.

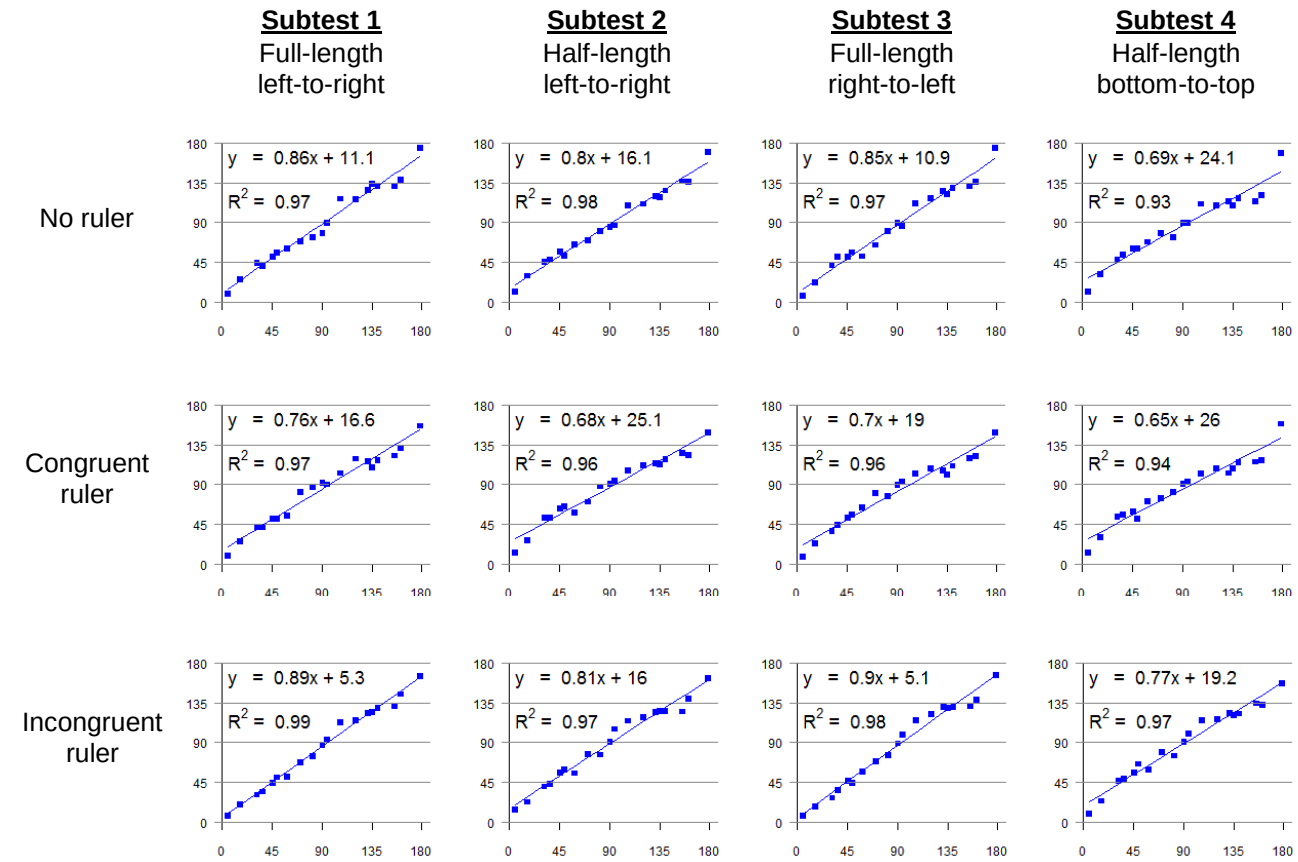
Appendix B-3

Mean PAE across all Training Blocks and Post-subtests – *Incongruent ruler* Condition



Plots of individual participants in the *incongruent ruler* condition show progression of mean errors through all training blocks (consisting of 8 trials each) and subtests (consisting of 19 trials each). Blue points indicate training blocks, while black squares indicate post-subtests.

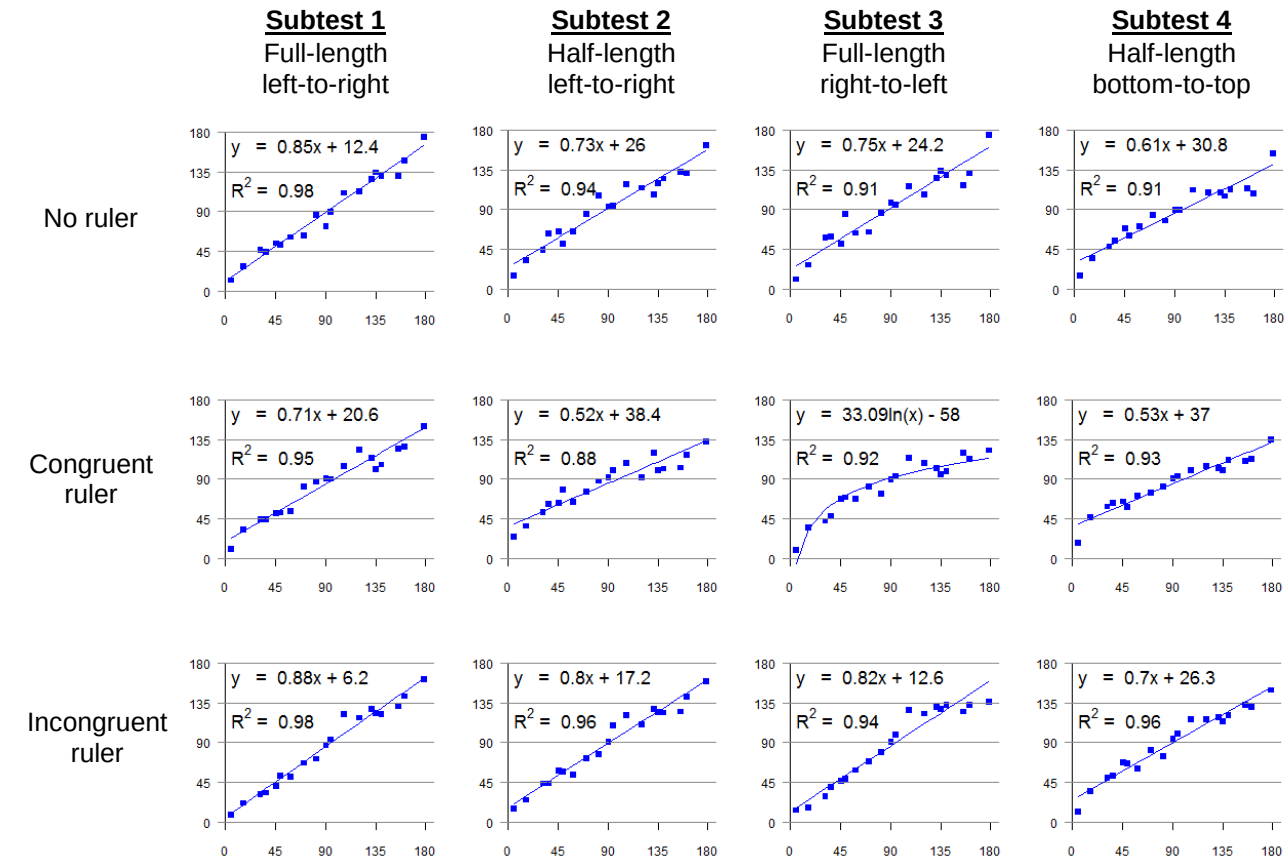
Appendix C-1
Median Estimates over Actual Magnitudes – All Participants



Median estimates of all participants. X-axis represents actual magnitude, Y-axis represents median estimated magnitude. Columns display progression of posttest blocks. Rows show different conditions.

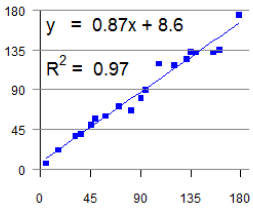
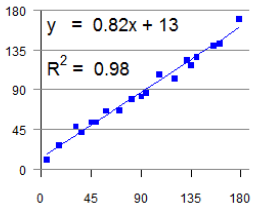
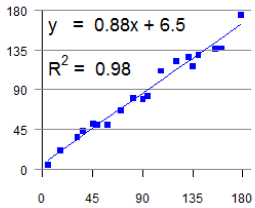
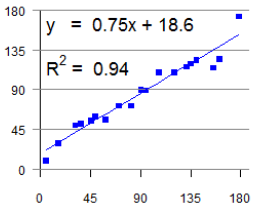
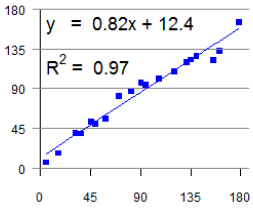
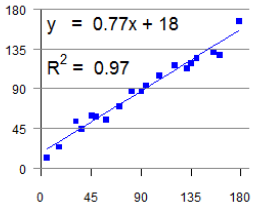
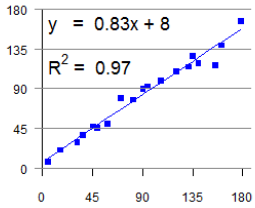
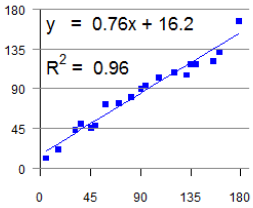
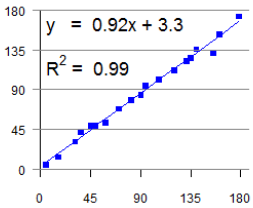
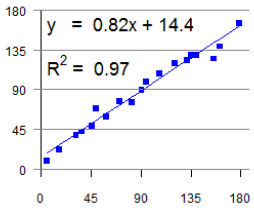
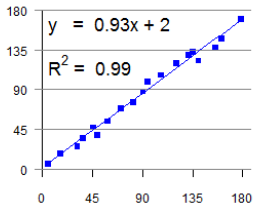
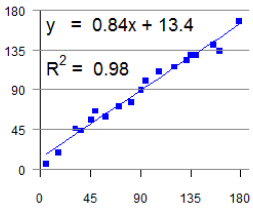
Appendix C-2

Median Estimates over Actual Magnitudes – Younger Children (Ages 7.3 – 8.66)



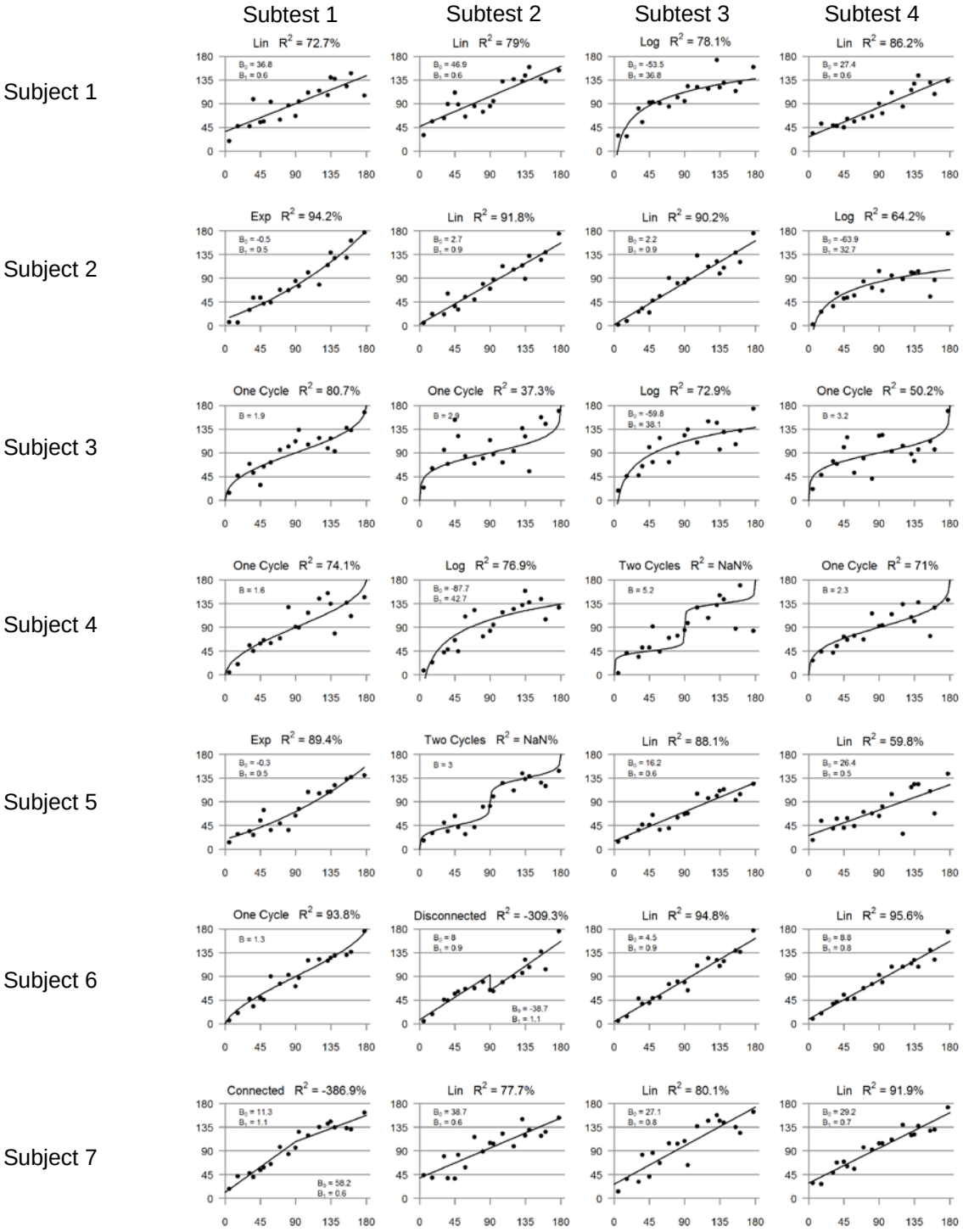
Median estimates of younger half of participants. X-axis represents actual magnitude, Y-axis represents median estimated magnitude. Columns display progression of posttest blocks. Rows show different conditions.

Appendix C-3
Median Estimates over Actual Magnitudes – Older children (Ages 8.67 – 10.5)

	Subtest 1 Full-length left-to-right	**Subtest 2** Half-length left-to-right	**Subtest 3** Full-length right-to-left	**Subtest 4** Half-length bottom-to-top
No ruler				
Congruent ruler				
Incongruent ruler				

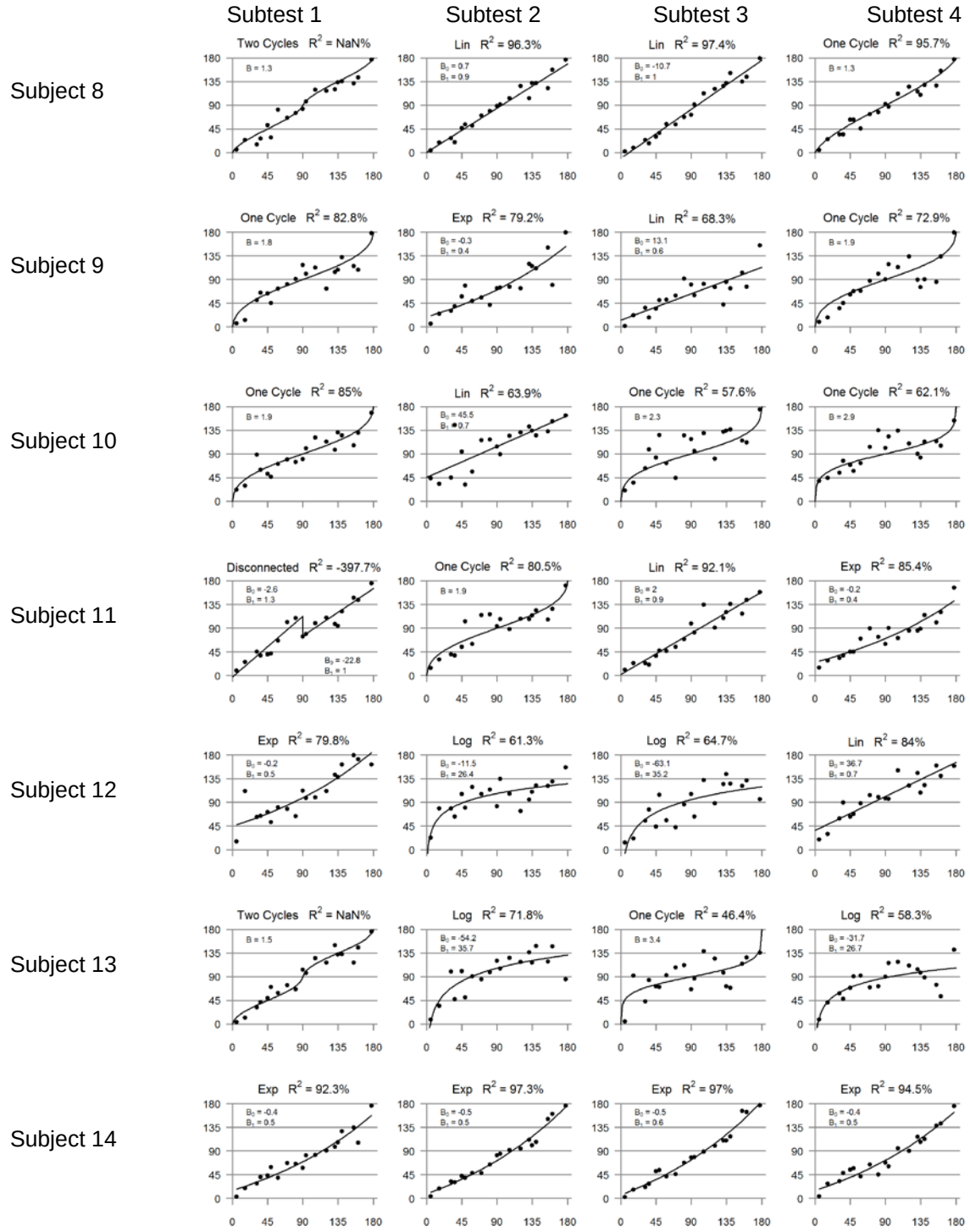
Median estimates of older half of participants. X-axis represents actual magnitude, Y-axis represents median estimated magnitude. Columns display progression of posttest blocks. Rows show different conditions.

Appendix D – 1.1
Individual Estimates over Actual Magnitudes – *No Ruler* Condition



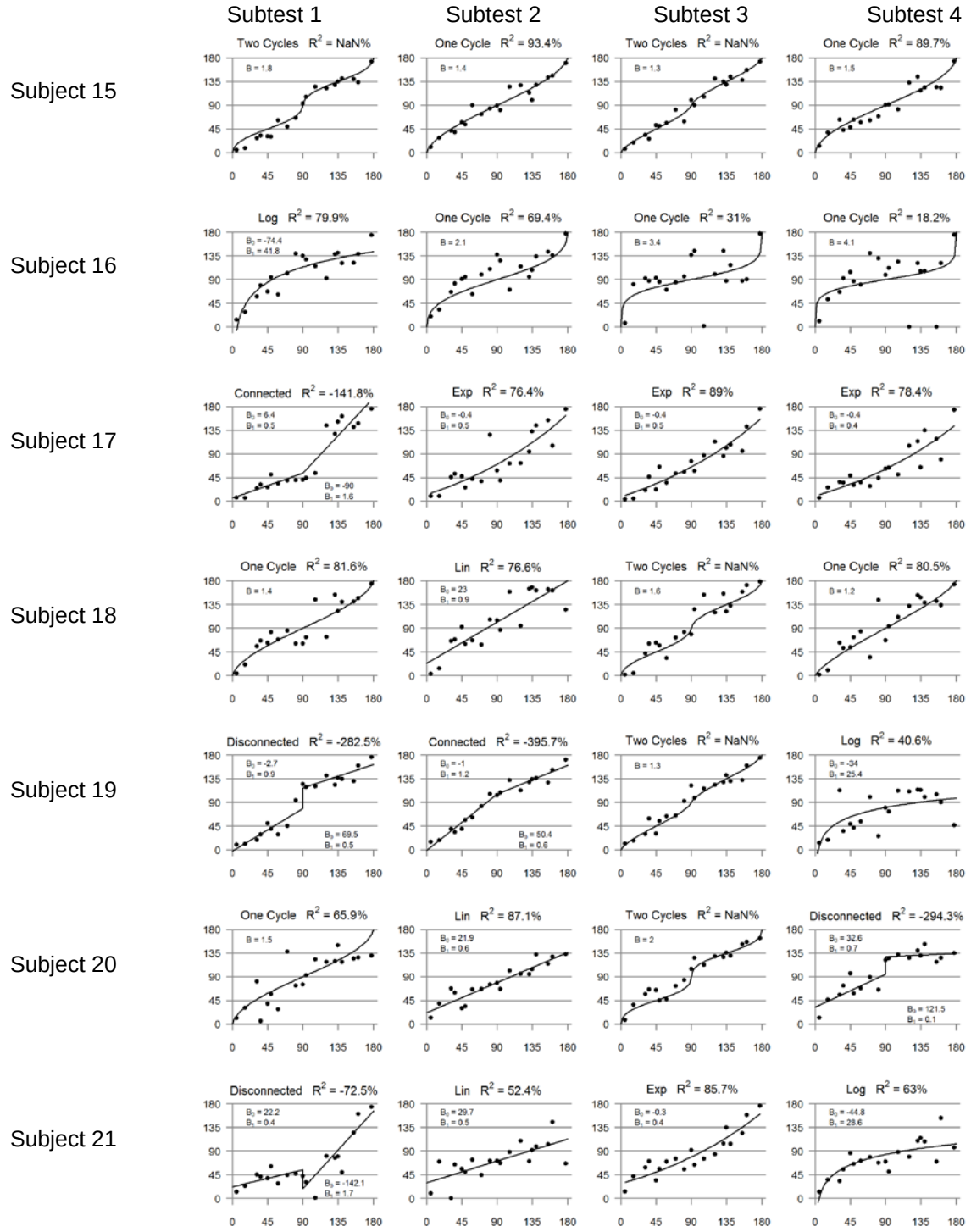
Description given on page 111.

Appendix D – 1.2
Individual Estimates over Actual Magnitudes – *No Ruler* Condition



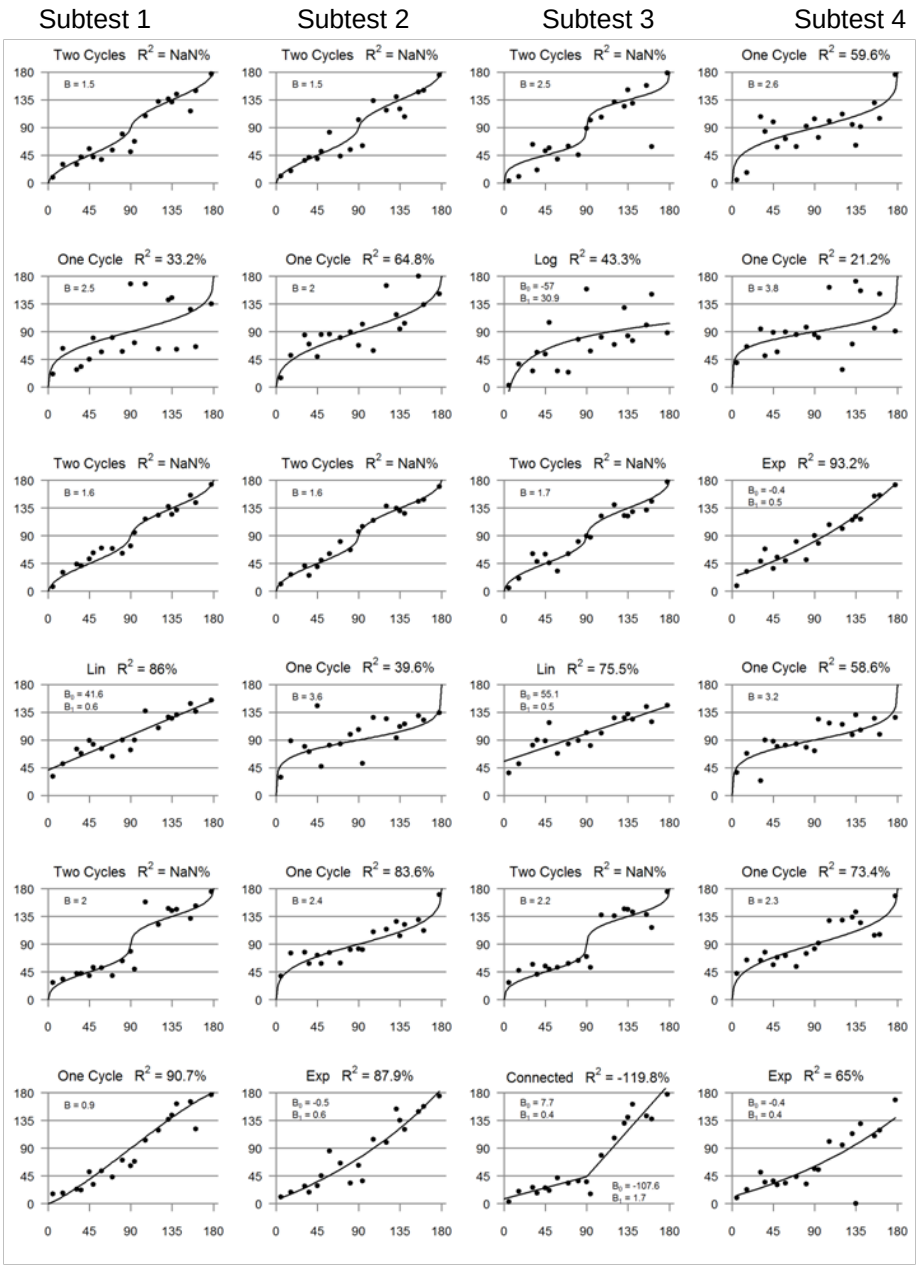
Description given on page 111.

Appendix D – 1.3
Individual Estimates over Actual Magnitudes – *No Ruler* Condition



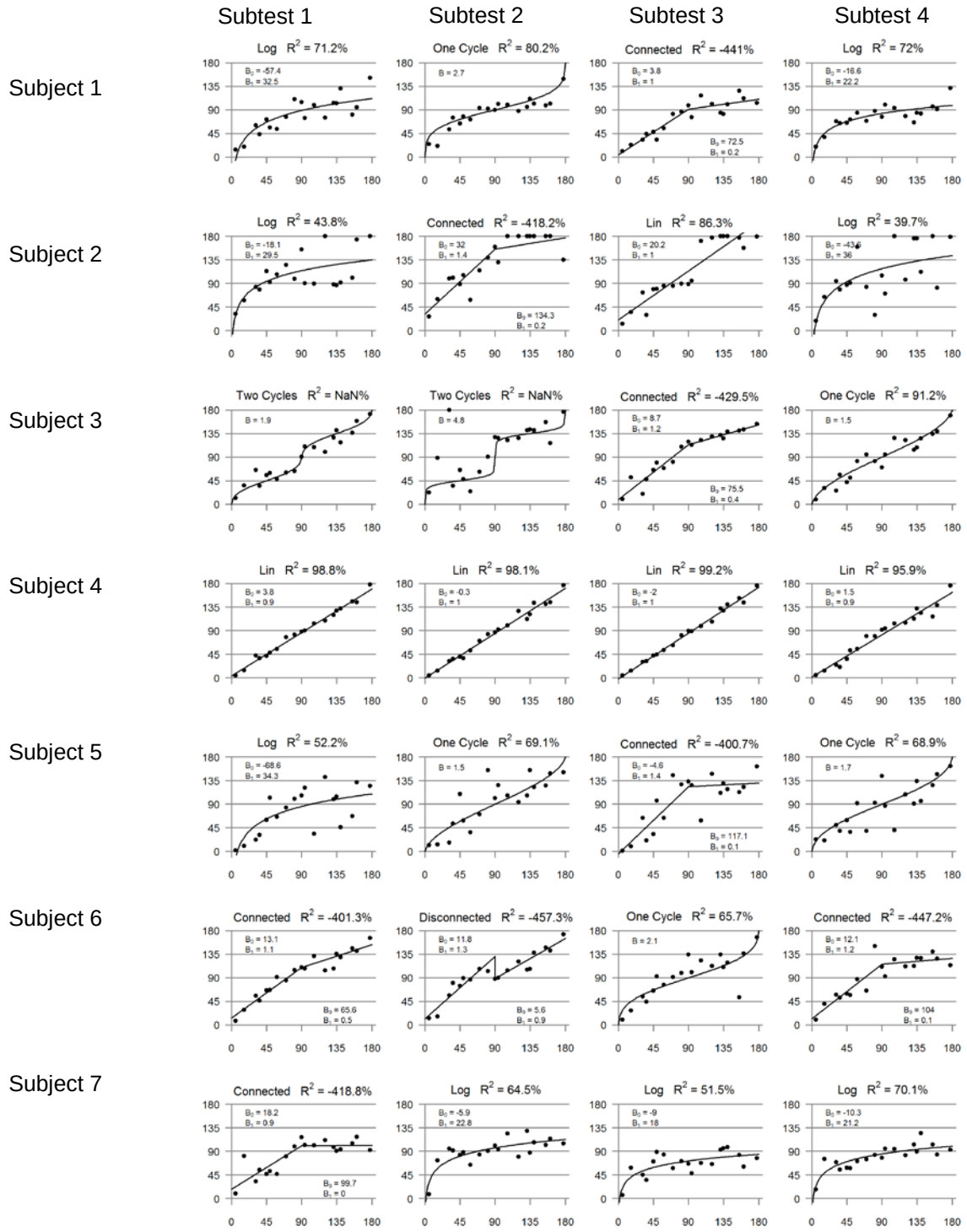
Description given on page 111.

Appendix D – 1.4
Individual Estimates over Actual Magnitudes – *No Ruler* Condition



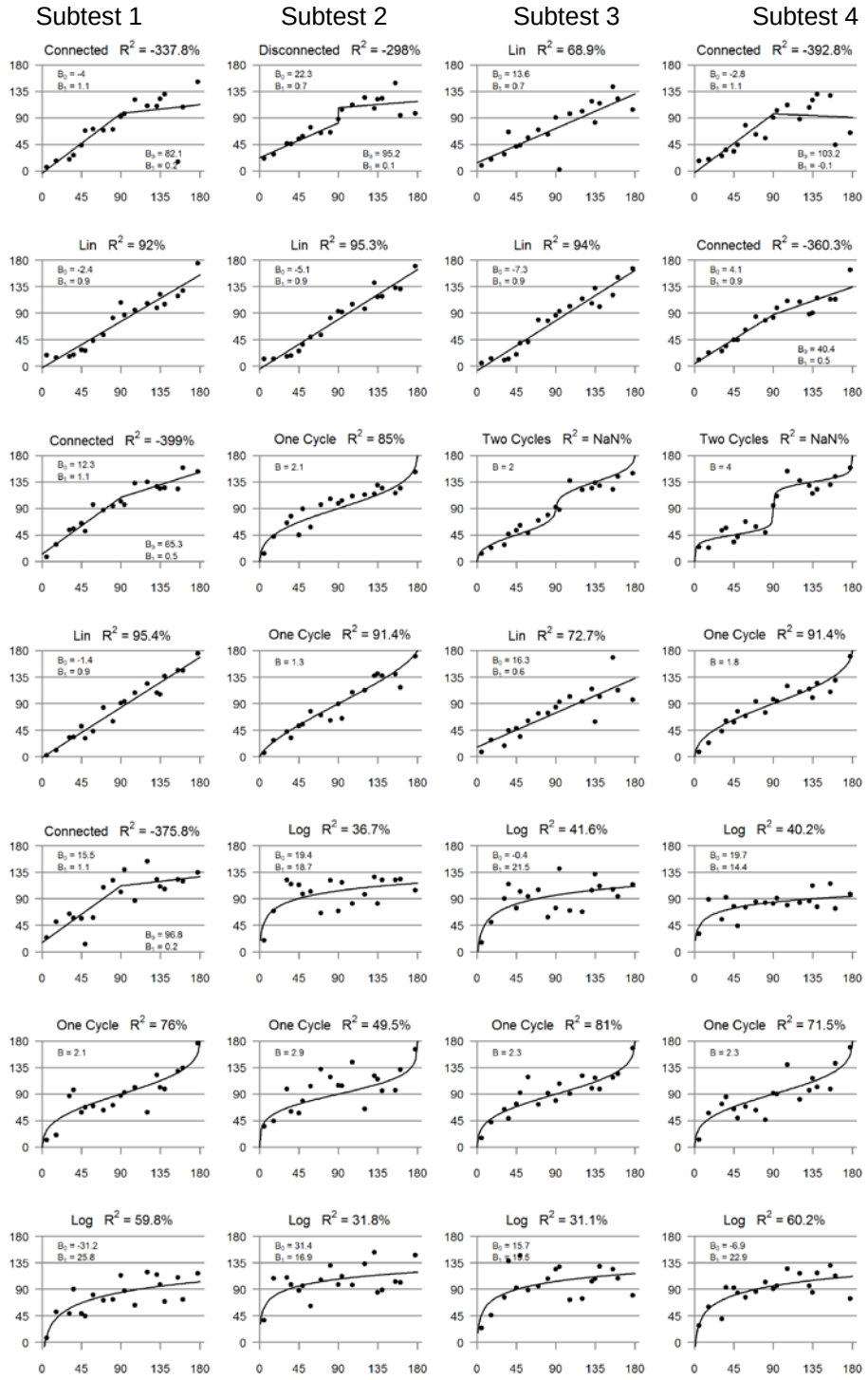
Description given on page 111.

Appendix D – 2.1
Individual Estimates over Actual Magnitudes – *Congruent Ruler* Condition



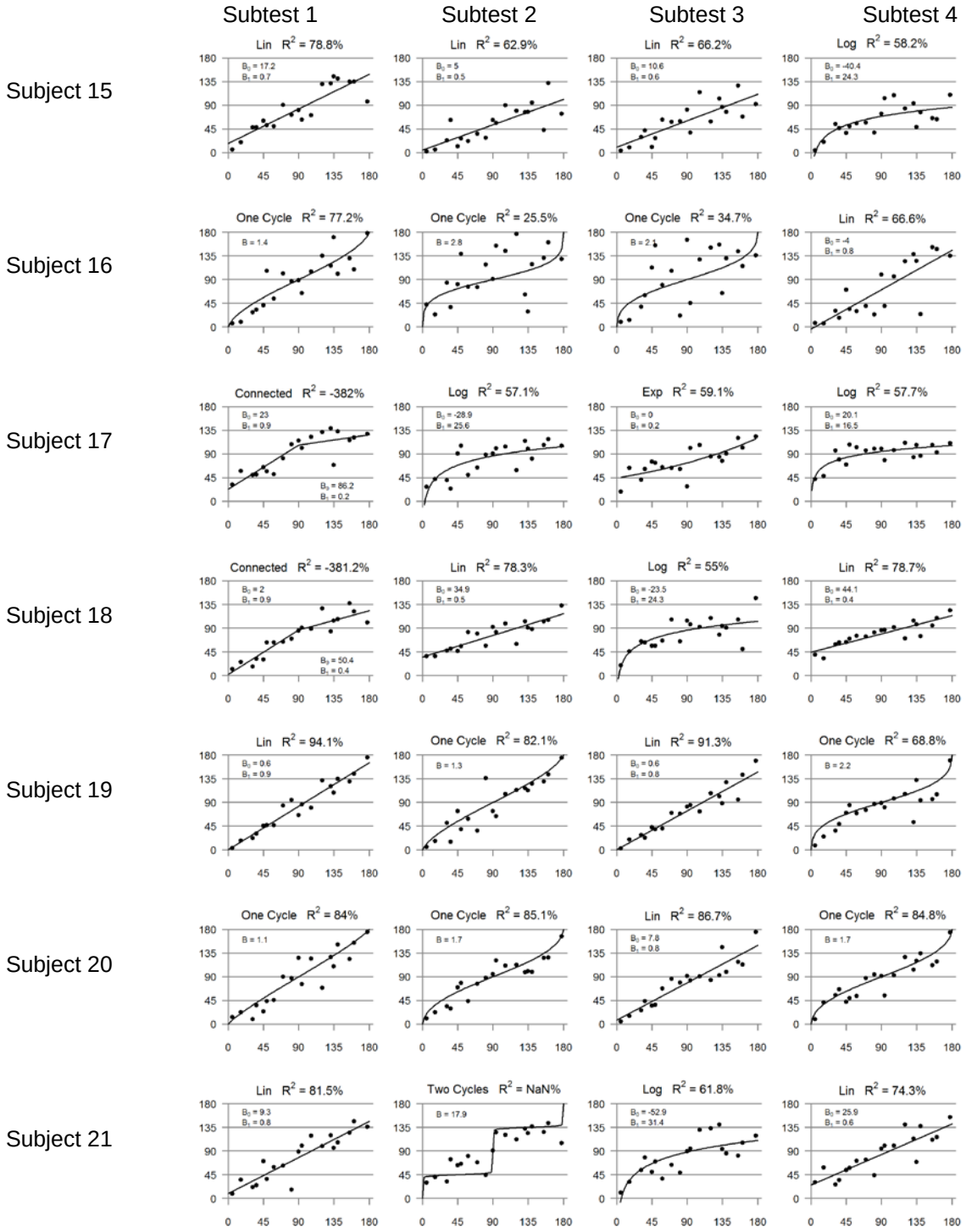
Description given on page 111.

Appendix D – 2.2
Individual Estimates over Actual Magnitudes – *Congruent Ruler* Condition



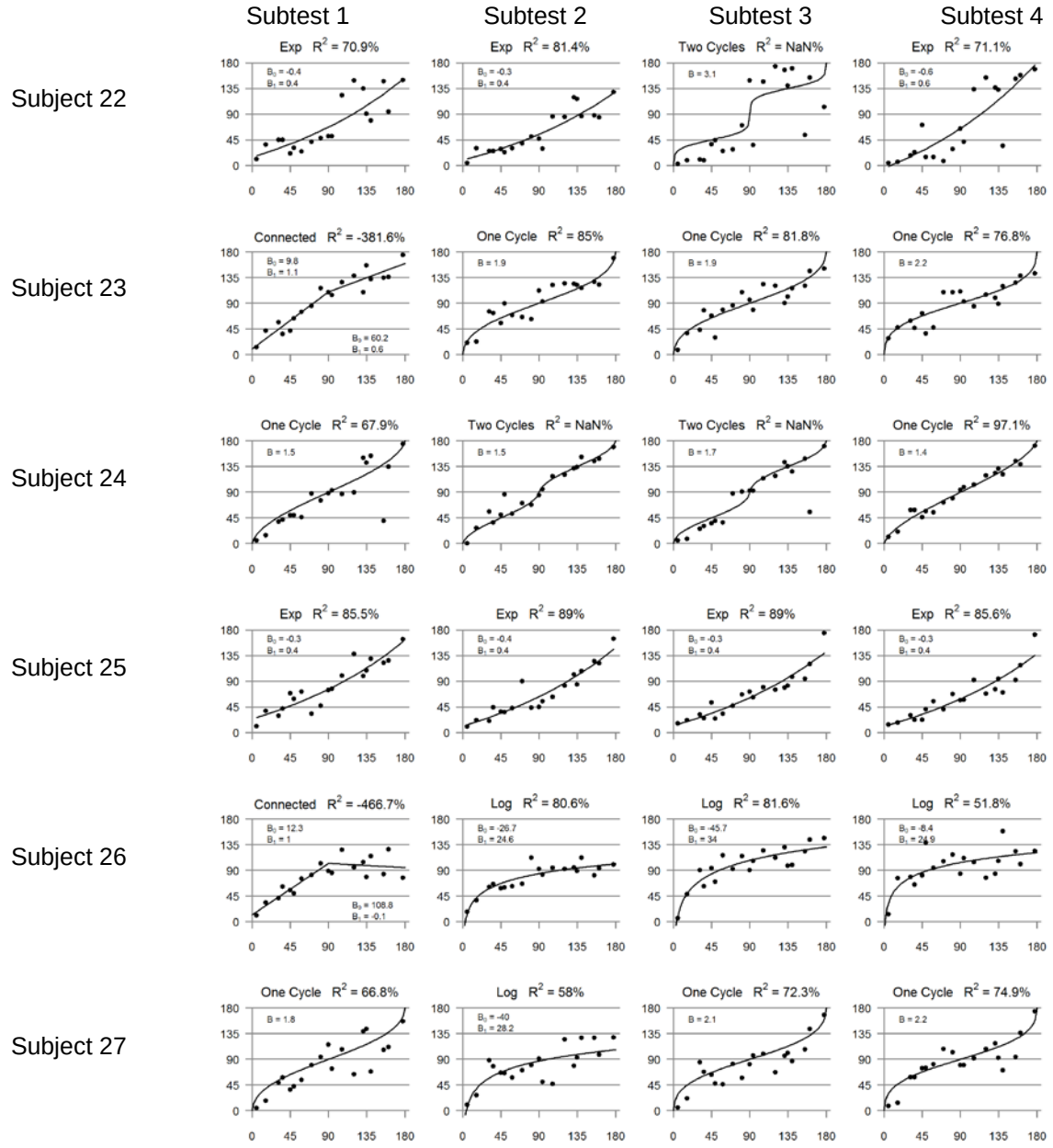
Description given on page 111.

Appendix D – 2.3
Individual Estimates over Actual Magnitudes – *Congruent Ruler* Condition



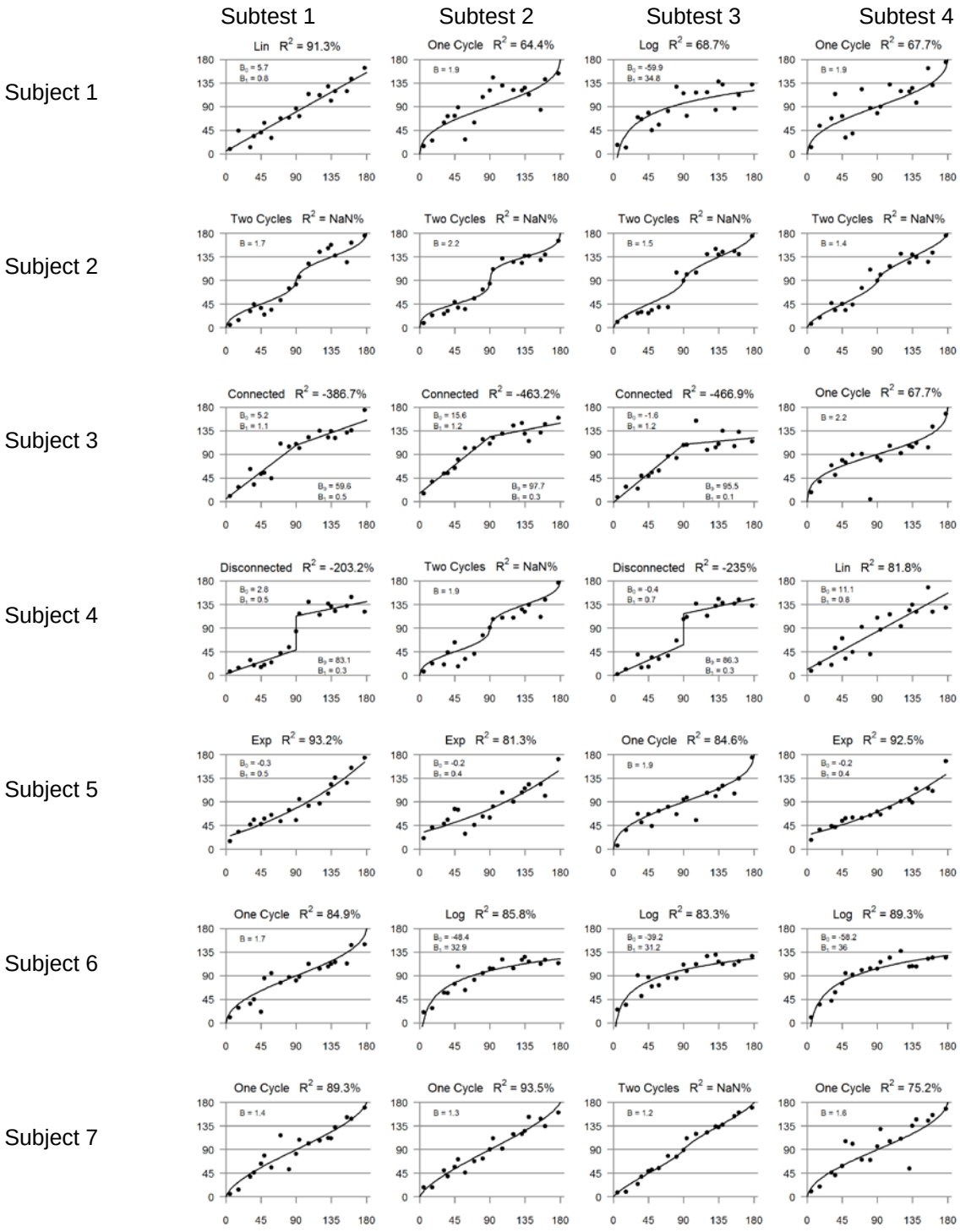
Description given on page 111.

Appendix D – 2.4
Individual Estimates over Actual Magnitudes – *Congruent Ruler* Condition



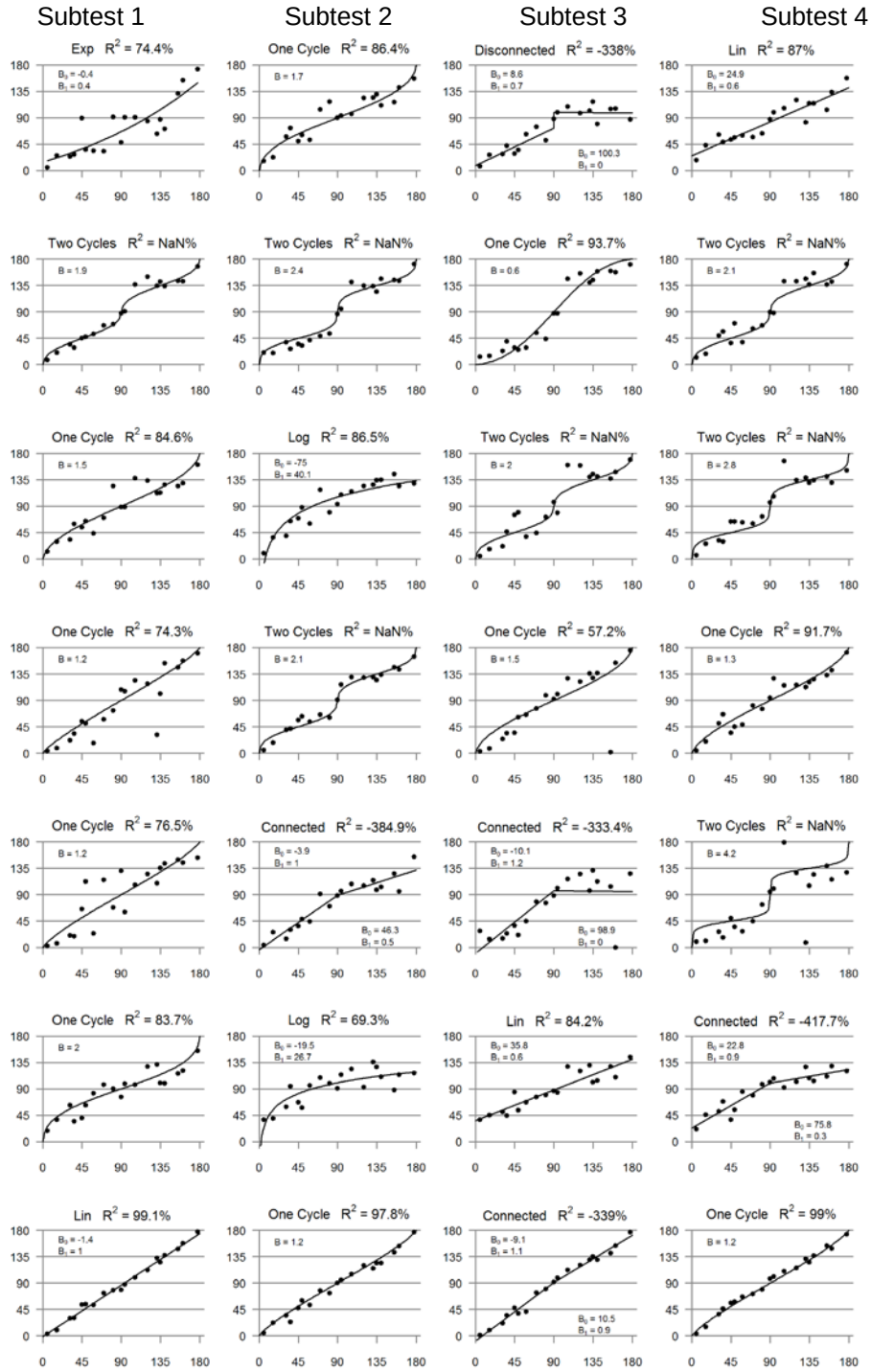
Description given on page 111.

Appendix D – 3.1
Individual Estimates over Actual Magnitudes – *Incongruent Ruler* Condition



Description given on page 111.

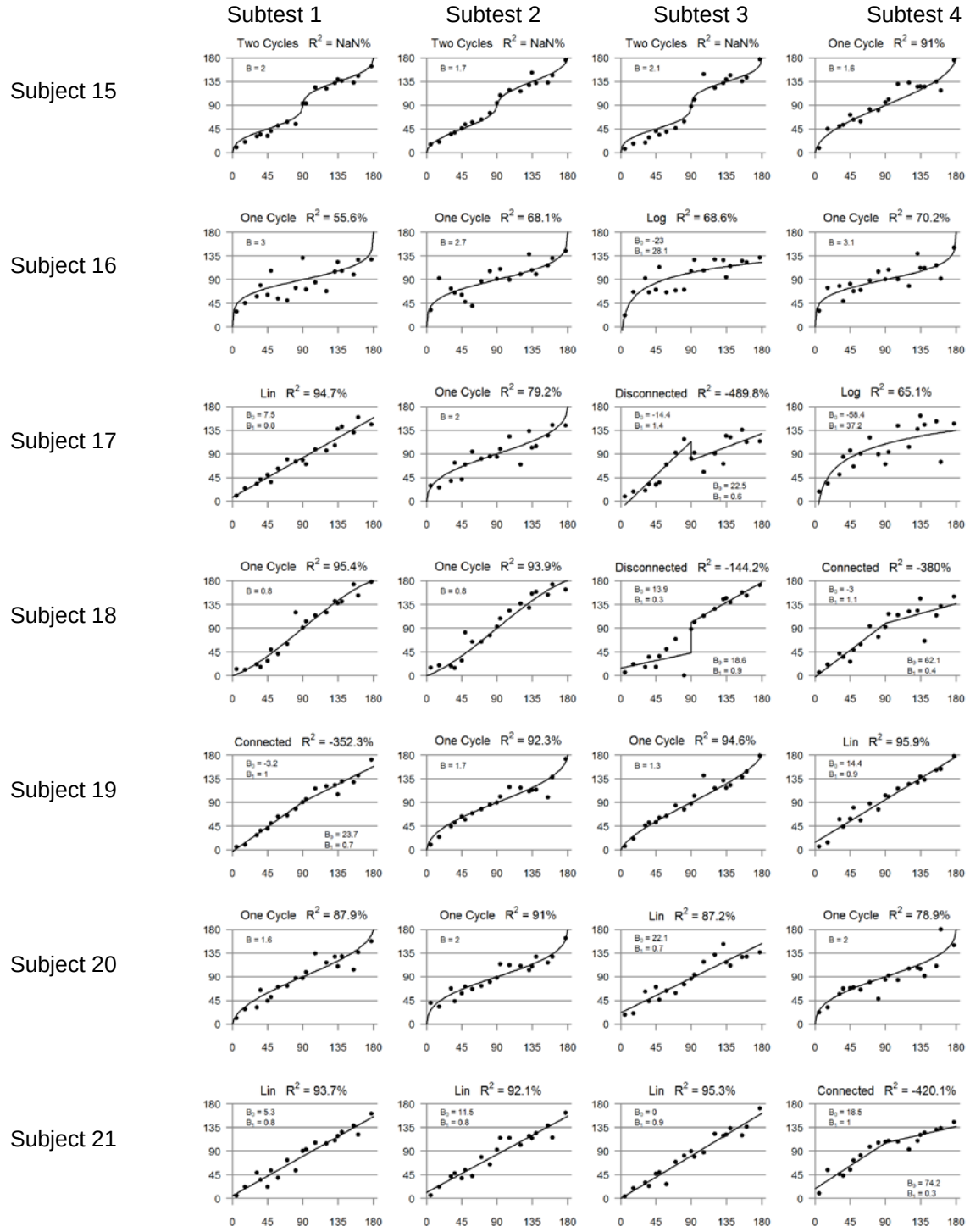
Appendix D – 3.2
Individual Estimates over Actual Magnitudes – *Incongruent Ruler* Condition



Description given on page 111.

Appendix D – 3.3

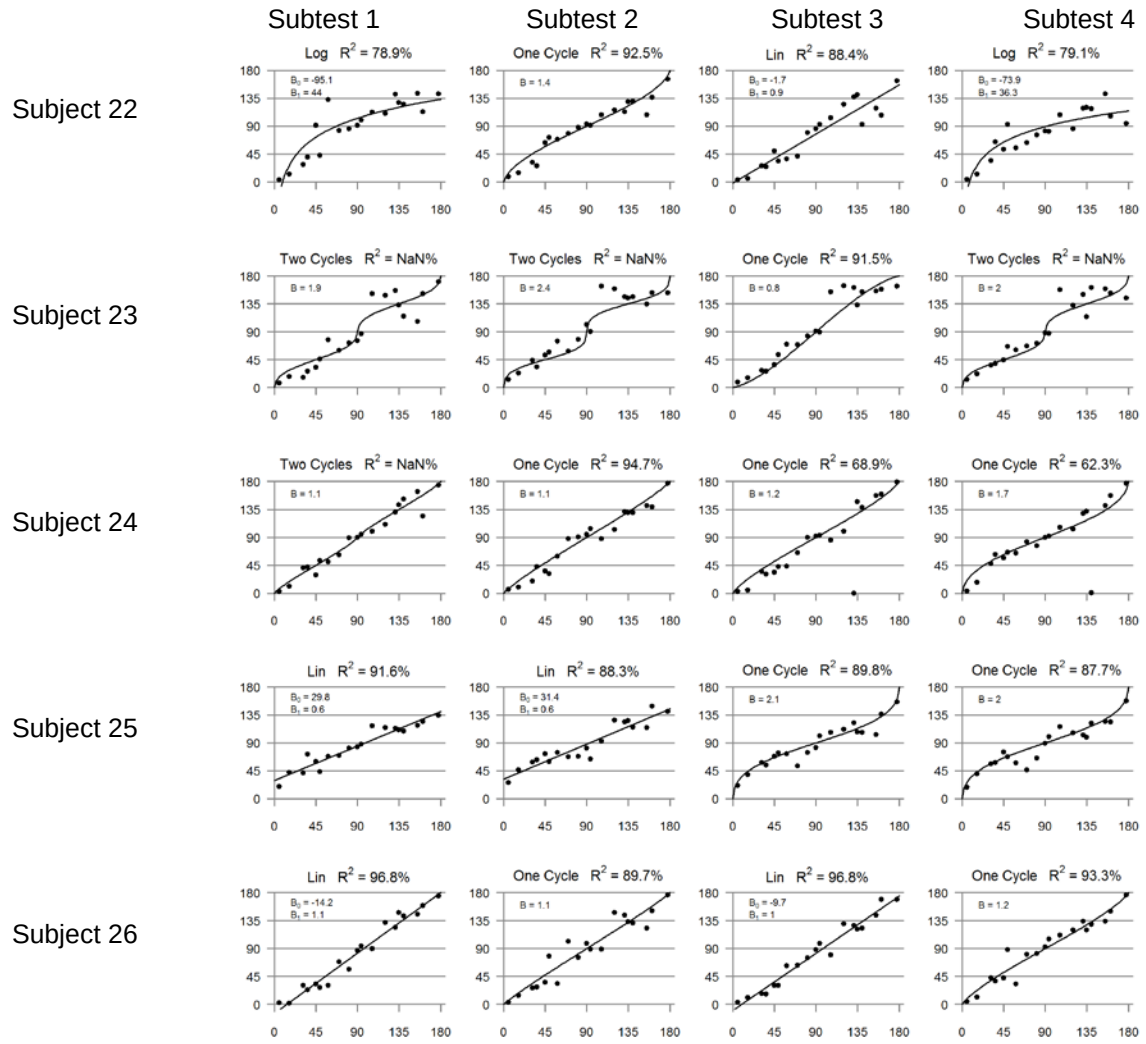
Individual Estimates over Actual Magnitudes – *Incongruent Ruler* Condition



Description given on page 111.

Appendix D – 3.4

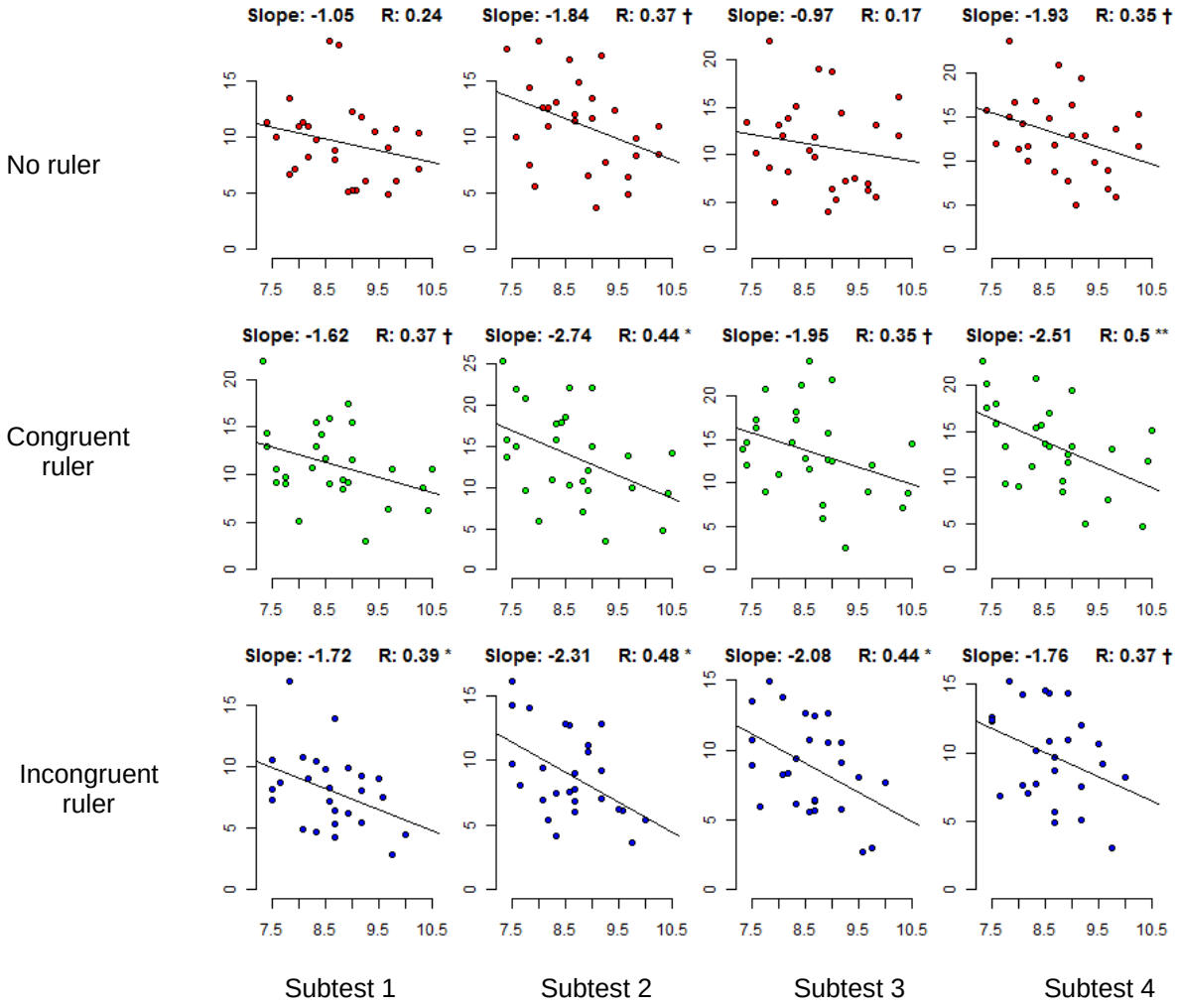
Individual Estimates over Actual Magnitudes – *Incongruent Ruler* Condition



For every individual participant four graphs corresponding to each subtest are shown depicted estimated magnitudes vs. actual magnitudes. For graphs in Appendix J, experiment 2, only three subtests are displayed. Each graph shows either the best fitting linear, logarithmic, exponential, one-cycle power model, two-cycle power model, connected segmented regression model (with a 90 breakpoint), or disconnected segmented regression model (with a 90 breakpoint). For each model the variance explained is displayed in the title of the graph. Estimated parameters are also displayed. In the case of segmented models the parameter(s) for the second segment are shown in the bottom-right corner. In the case of the exponential function the parameters displayed were generated out of a scaled data set (0 – 1, instead of 0 – 180).

Appendix E - 1

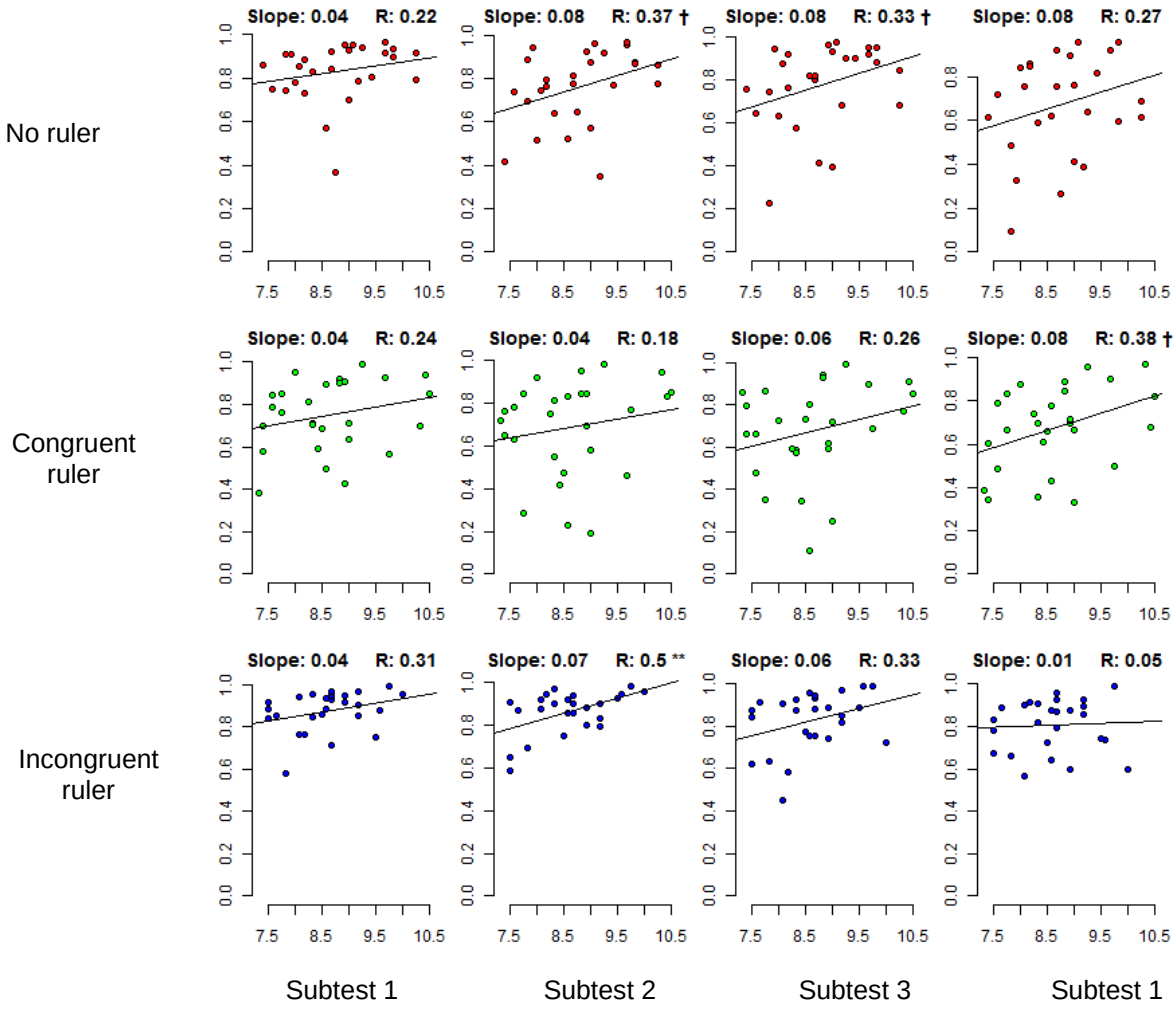
Mean PAE vs. age across all Subtests, by Condition



Displays the relationship between Mean PAE and age (in months) across all four subtests. Title of each graph includes slope of the regression line and the correlation between mean PAE and age.

Appendix E – 2

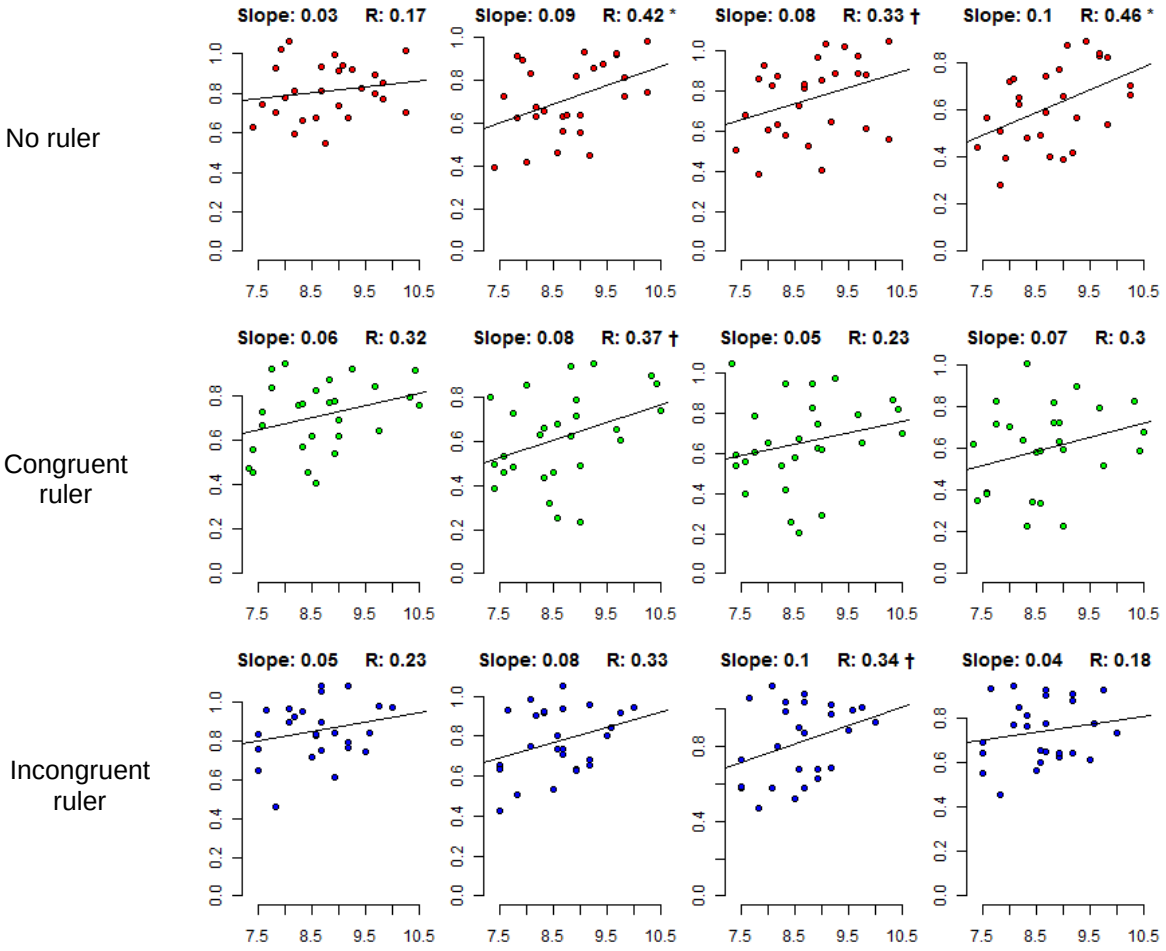
Linearity vs. Age across all Subtests, by Condition



Displays the relationship between Linearity and age (in months) across all four subtests. Title of each graph includes slope of the regression line and the correlation between Linearity and age.

Appendix E – 3

Slope vs. Age across all Subtests, by Condition



Displays the relationship between Slope and age (in months) across all four subtests. Title of each graph includes slope of the regression line and the correlation between Slope and age.

Appendix F – Experiment 1 Posttest ANOVAs and ANCOVAs

ANOVA – Mean PAE

Source	SS	df	MS	F	η_p^2
Between					
Condition	.089	2	.044	8.32 **	.18
Congruent ruler vs. others	.06	1	.062	11.6 **	.12
No ruler vs. Incongruent ruler	.026	1	.026	4.9 *	.05
Error	.411	77	.005		
Within					
Subtest	.027	3	.009	15.91***	.17
Subtest × Condition	.006	6	.001	1.73	.04
Error	.132	231	.001		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Mean PAE with Age (in months)

Source	SS	df	MS	F	η_p^2
Between					
Condition	.103	2	.051	12.22 ***	.24
Congruent ruler vs. others	.064	1	.064	16.2 **	.17
No ruler vs. Incongruent ruler	.040	1	.040	9.5 **	.11
Age (# months)	.091	1	.091	21.62 ***	.22
Error	.320	76	.004		
Within					
Subtest	.027	3	.009	16.1 ***	.18
Subtest × Condition	.006	6	.001	1.87 †	.05
Subtest × Age (in months)	.003	3	.001	1.91	.03
Error	.129	228	.001		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA –Linearity

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.106	2	.553	7.3 **	.16
Congruent ruler vs. others	.75	1	.76	9.9 **	.11
No ruler vs. Incongruent ruler	.35	1	.35	4.6 *	.06
Error	5.833	77	.076		
Within					
Subtest	.395	3	.132	9.39***	.11
Subtest × Condition	.107	6	.018	1.28	.03
Error	3.237	231	.014		
		319			

***** p < .001 ** p < .01 * p < .05 † p < .1**

ANOVA – Linearity with Age (in months)

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.251	2	.626	9.42 ***	.20
Congruent ruler vs. others	.77	1	.77	11.6 **	.13
No ruler vs. Incongruent ruler	.49	1	.49	6.4 **	.09
Age (# months)	.79	1	.786	11.8 **	.14
Error	5.047	76	.066		
Within					
Subtest	.395	3	.13	9.32	.11
Subtest × Condition	.11	6	.02	1.32	.03
Subtest × Age (# months)	.02	3	.01	.42	.01
Error	3.22	228	.01		
	319				

*** p < .001 ** p < .01 ‘ * p < .05 † p < .1

ANOVA – Slope

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.236	2	.618	6.3 **	.14
Congruent ruler vs. others	.96	1	.76	9.8 **	.11
No ruler vs. Incongruent ruler	.26	1	.35	2.7	.03
Error	7.542	77	.098		
Within					
Subtest	.862	3	.287	29.0***	.27
Subtest × Condition	.089	6	.015	1.5	.04
Error	2.288	231	.010		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Slope with Age (in months)

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.40	2	.70	8.4 **	.18
Congruent ruler vs. others	.99	1	.99	11.8 **	.13
No ruler vs. Incongruent ruler	.42	1	.42	5.0 *	.06
Age (# months)	1.16	1	1.16	13.8 ***	.15
Error	6.38	76	.084		
Within					
Subtest	.86	3	.287	29.1 ***	.28
Subtest × Condition	.09	6	.015	1.6	.04
Subtest × Age (# months)	.04	3	.014	1.4	.02
Error	2.25	228	.010		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Slope with 178 removed

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.33	2	.663	6.9 **	.15
Congruent ruler vs. others	.90	1	.90	9.3 **	.11
No ruler vs. Incongruent ruler	.41	1	.41	4.3 *	.05
Error	7.43	77	.097		
Within					
Subtest	1.00	3	.334	25.9***	.25
Subtest × Condition	.110	6	.018	1.4	.04
Error	2.989	231	.013		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Slope with Age (in months) with 178 removed

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.48	2	.74	8.4 ***	.18
Congruent ruler vs. others	.93	1	.99	10.6 **	.12
No ruler vs. Incongruent ruler	.57	1	.42	6.5 *	.08
Age (# months)	.77	1	.77	8.7 **	.10
Error	6.667	76	.088		
Within					
Subtest	1.00	3	.33	25.8 ***	.25
Subtest × Condition	.11	6	.019	1.5	.04
Subtest × Age (# months)	.04	3	.013	1.0	.01
Error	2.95	228	.013		
		319			

***** p < .001 ** p < .01 * p < .05 † p < .1**

ANOVA – Slope with 5 and 178 removed

Source	SS	df	MS	F	η_p^2
Between					
Condition	1.73	2	.87	6.8 **	.15
Congruent ruler vs. others	1.17	1	1.17	9.3 **	.11
No ruler vs. Incongruent ruler	.54	1	.54	4.3 *	.05
Error	9.82	77	.13		
Within					
Subtest	1.27	3	.422	24.6 ***	.24
Subtest × Condition	.142	6	.024	1.4	.04
Error	3.96	231	.017		
		319			

*** p < .001 ** p < .01 * p < .05 † p < .1

ANOVA – Slope with Age (in months) with 5 and 178 removed

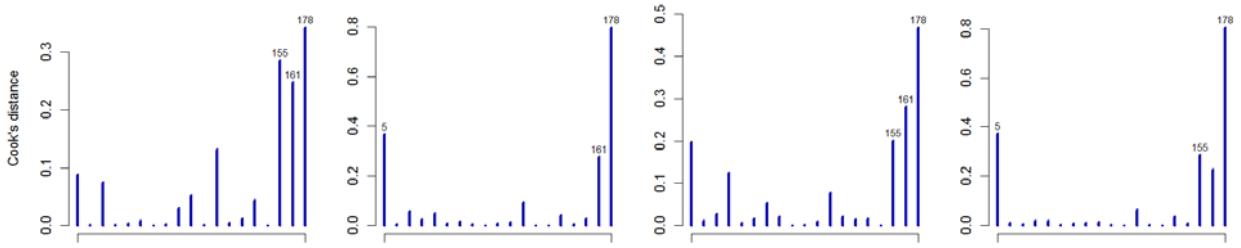
Source	SS	df	MS	F	η_p^2
Between					
Condition	1.93	2	.97	8.3 **	.18
Congruent ruler vs. others	1.21	1	.99	10.4 **	.12
No ruler vs. Incongruent ruler	.75	1	.42	6.7 *	.08
Age (# months)	1.00	1	1.00	8.6 **	.10
Error	8.82	76	.116		
Within					
Subtest	1.27	3	.42	24.6 ***	.25
Subtest × Condition	.15	6	.024	1.4	.04
Subtest × Age (# months)	.05	3	.017	1.0	.01
Error	3.912	228	.017		
		319			

***** p < .001 ** p < .01 * p < .05 † p < .1**

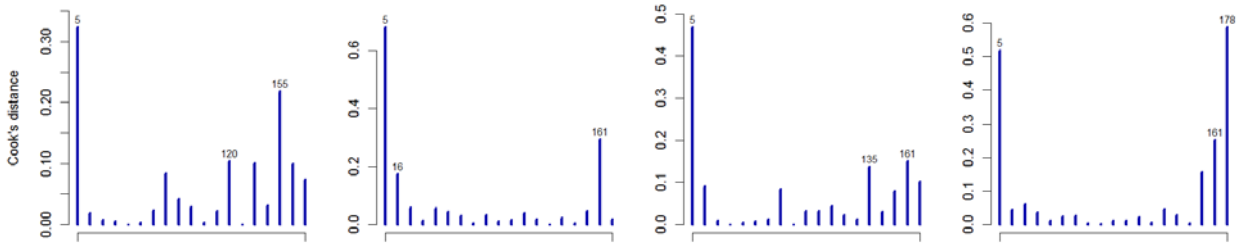
Appendix G

Cook's d for median target estimates across all subtests, by condition

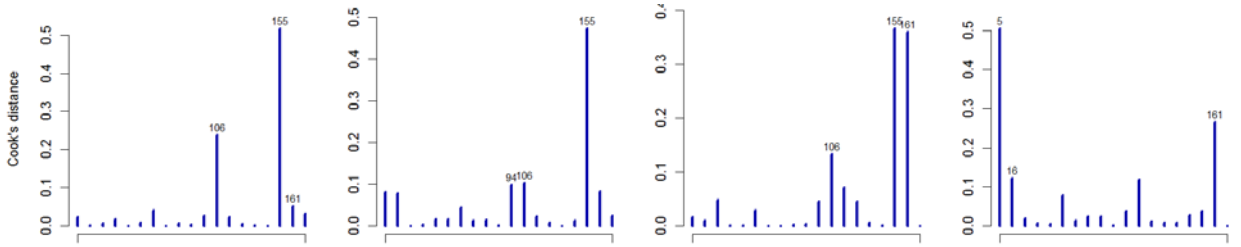
No ruler condition



Congruent ruler condition



Incongruent ruler condition

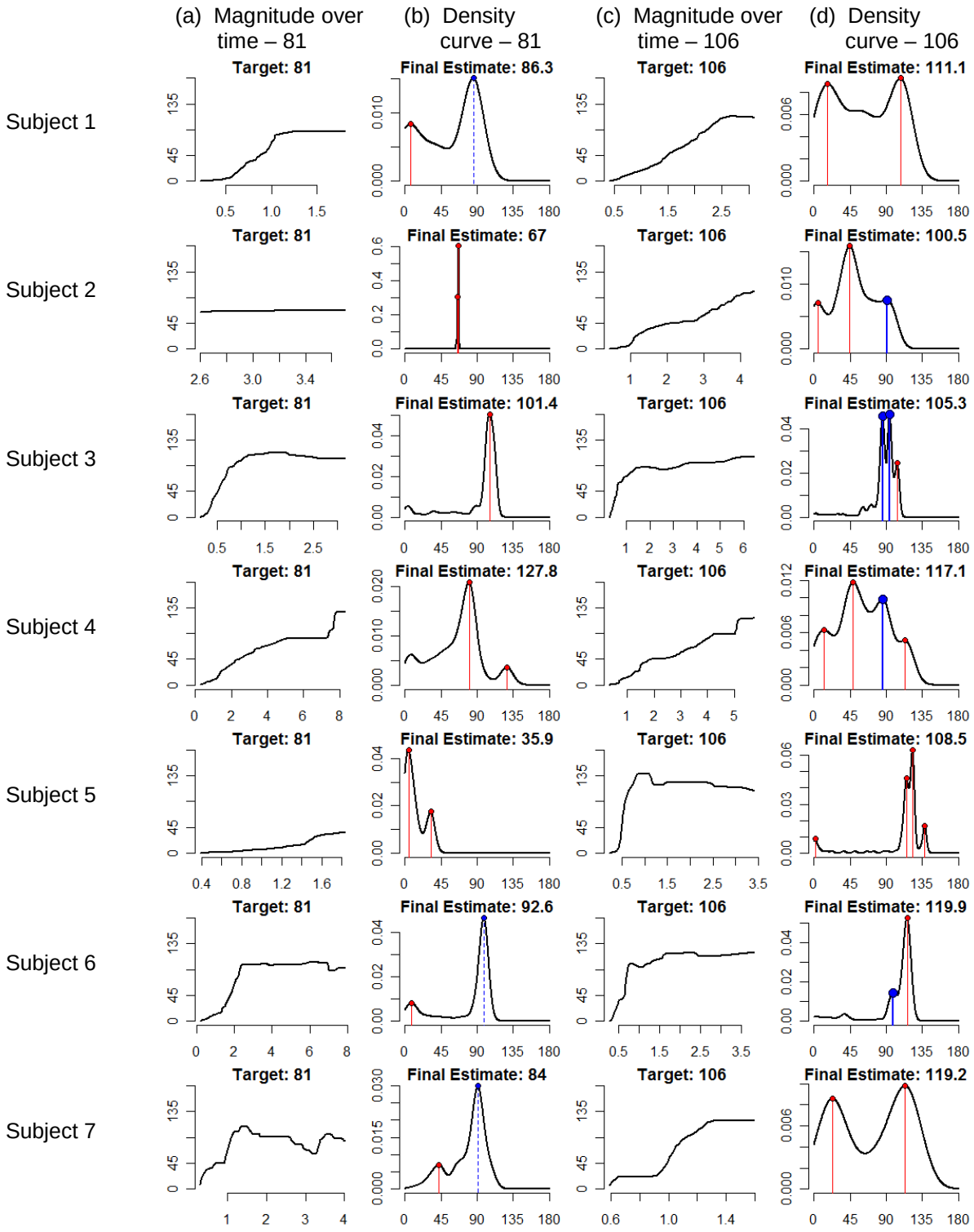


Subtest 1 Subtest 2 Subtest 3 Subtest 1

Each plot displays the Cook's d values for the median estimates at each target (19 in total). For the *no ruler* condition 178 has the largest Cook's d at each subtest. For the *congruent ruler* condition 5 has the largest or second largest Cook's d. For the *incongruent ruler* condition 155 and/or 161 have the largest Cook's d.

Appendix H – 1.1

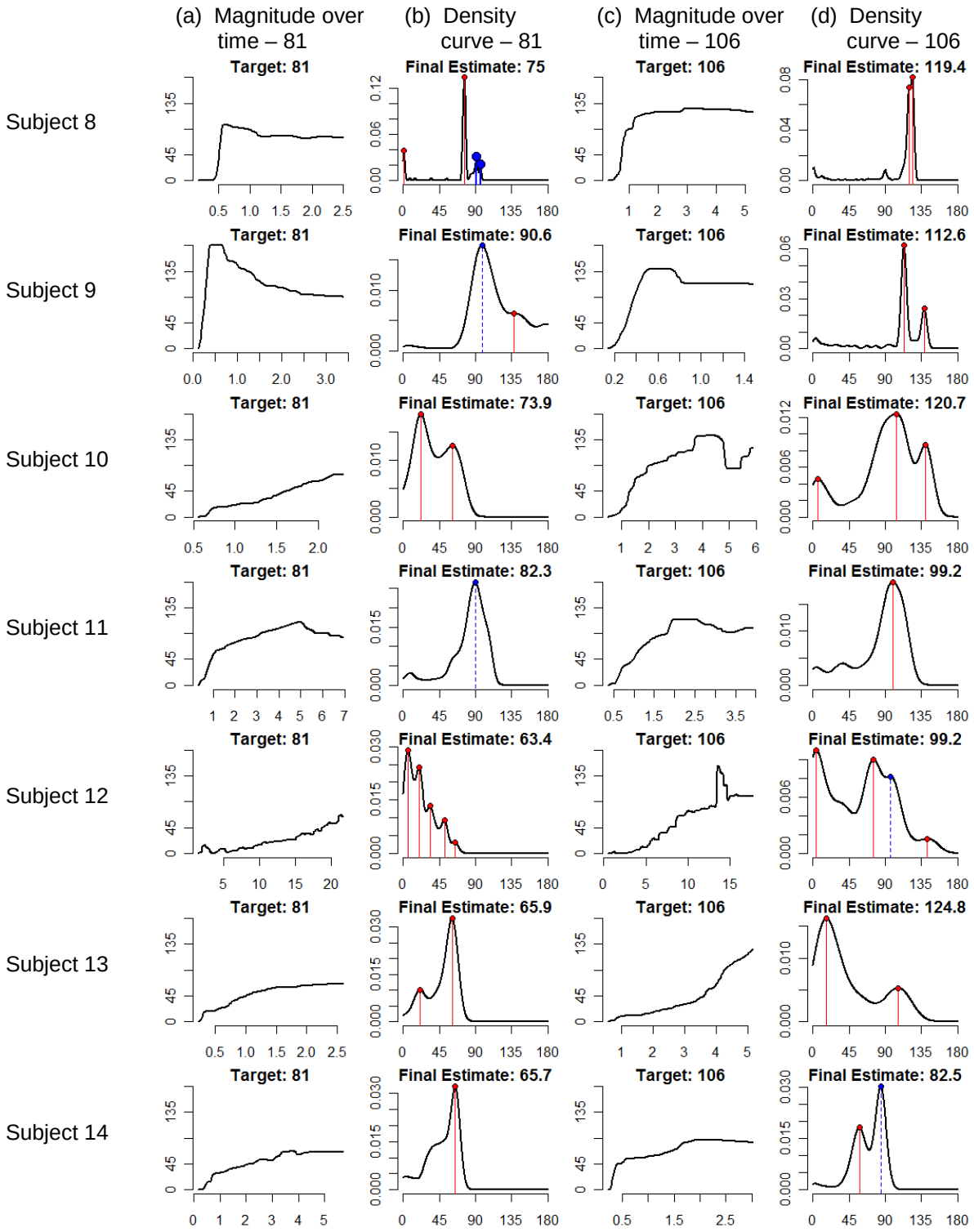
Trial curves and density plots (targets 81 and 106, subtest 1) – *No ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 1.2

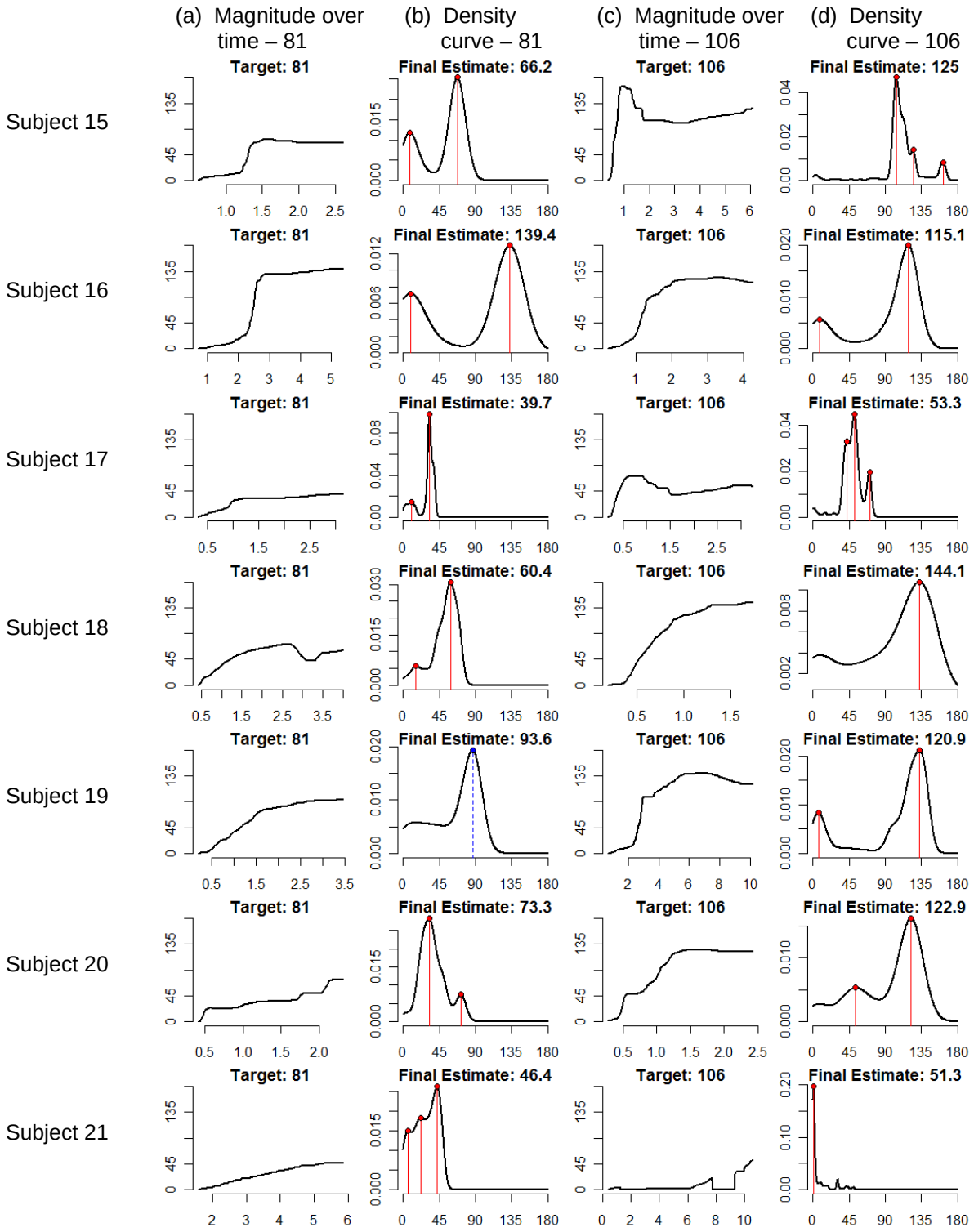
Trial curves and density plots (targets 81 and 106, subtest 1) – *No ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 1.3

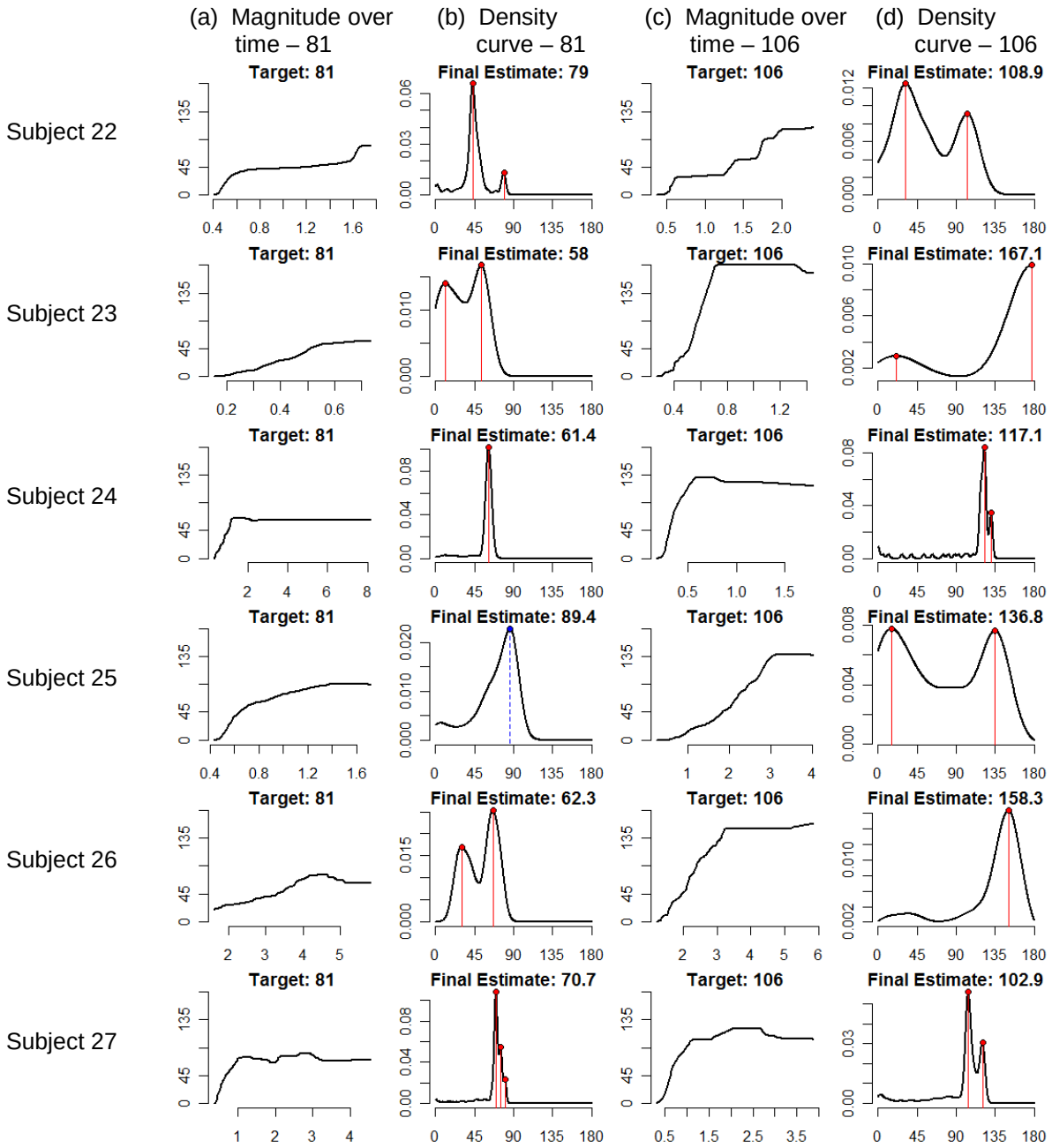
Trial curves and density plots (targets 81 and 106, subtest 1) – *No ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 1.4

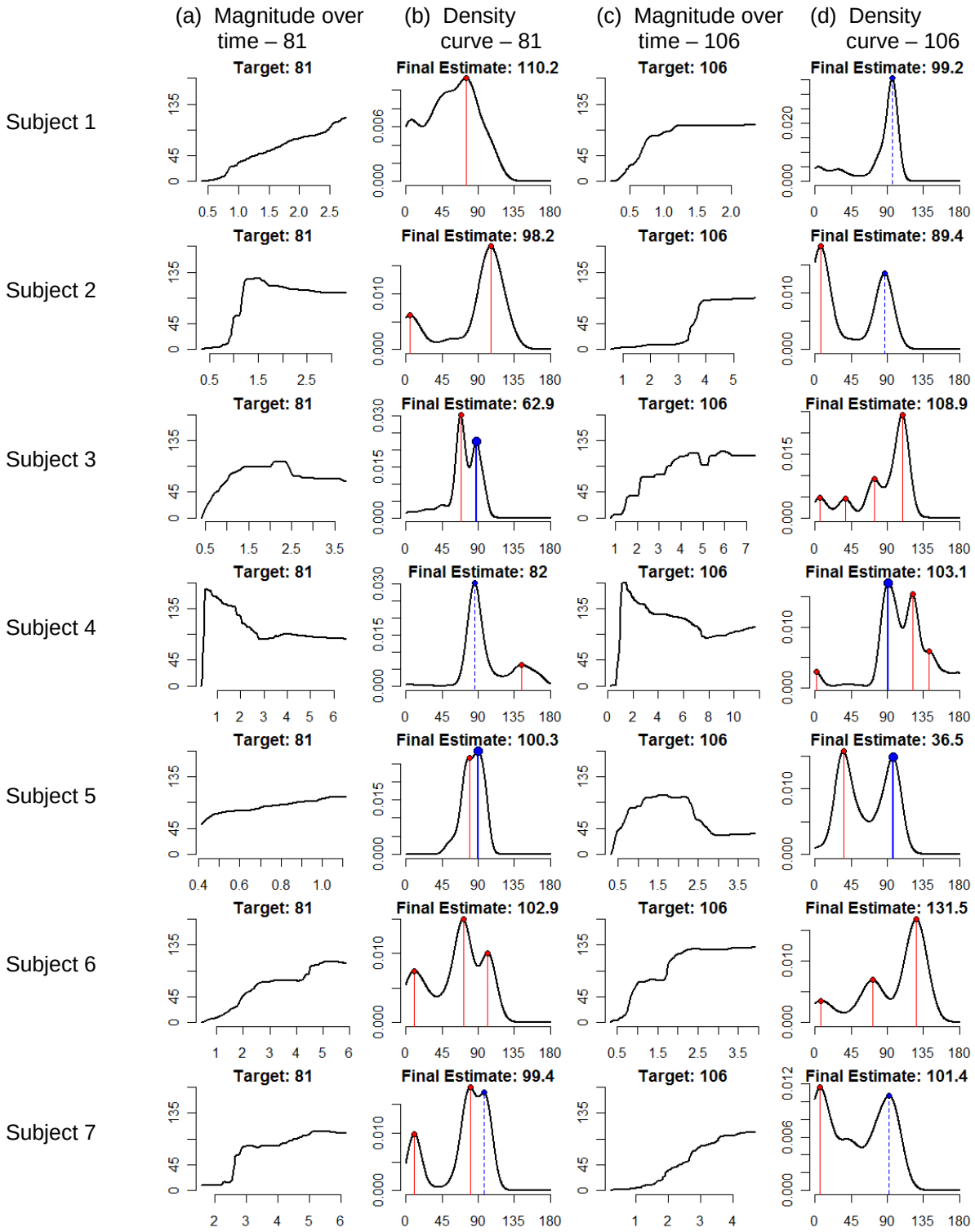
Trial curves and density plots (targets 81 and 106, subtest 1) – *No ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 2.1

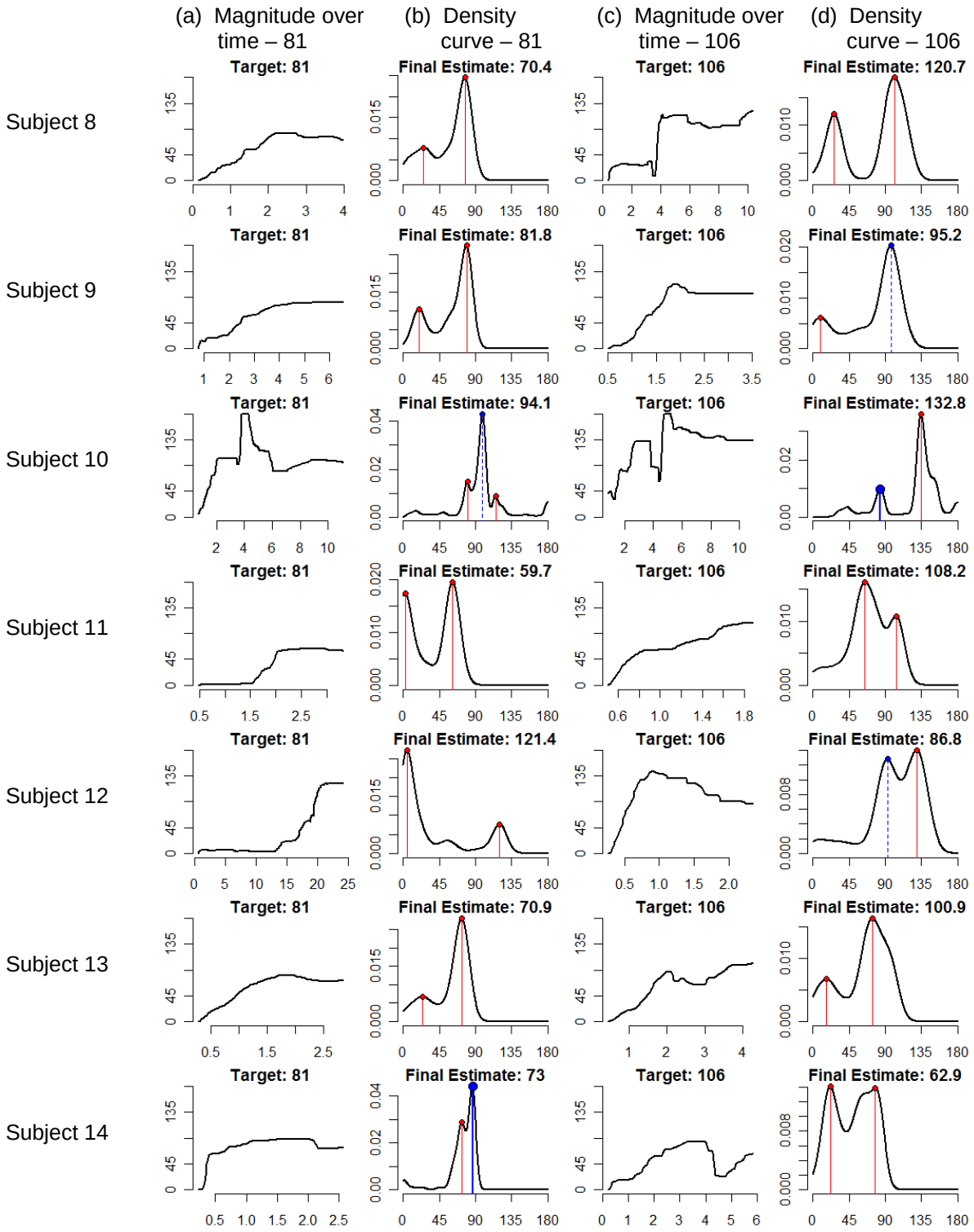
Trial curves and density plots (targets 81 and 106, subtest 1) – *Congruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 2.2

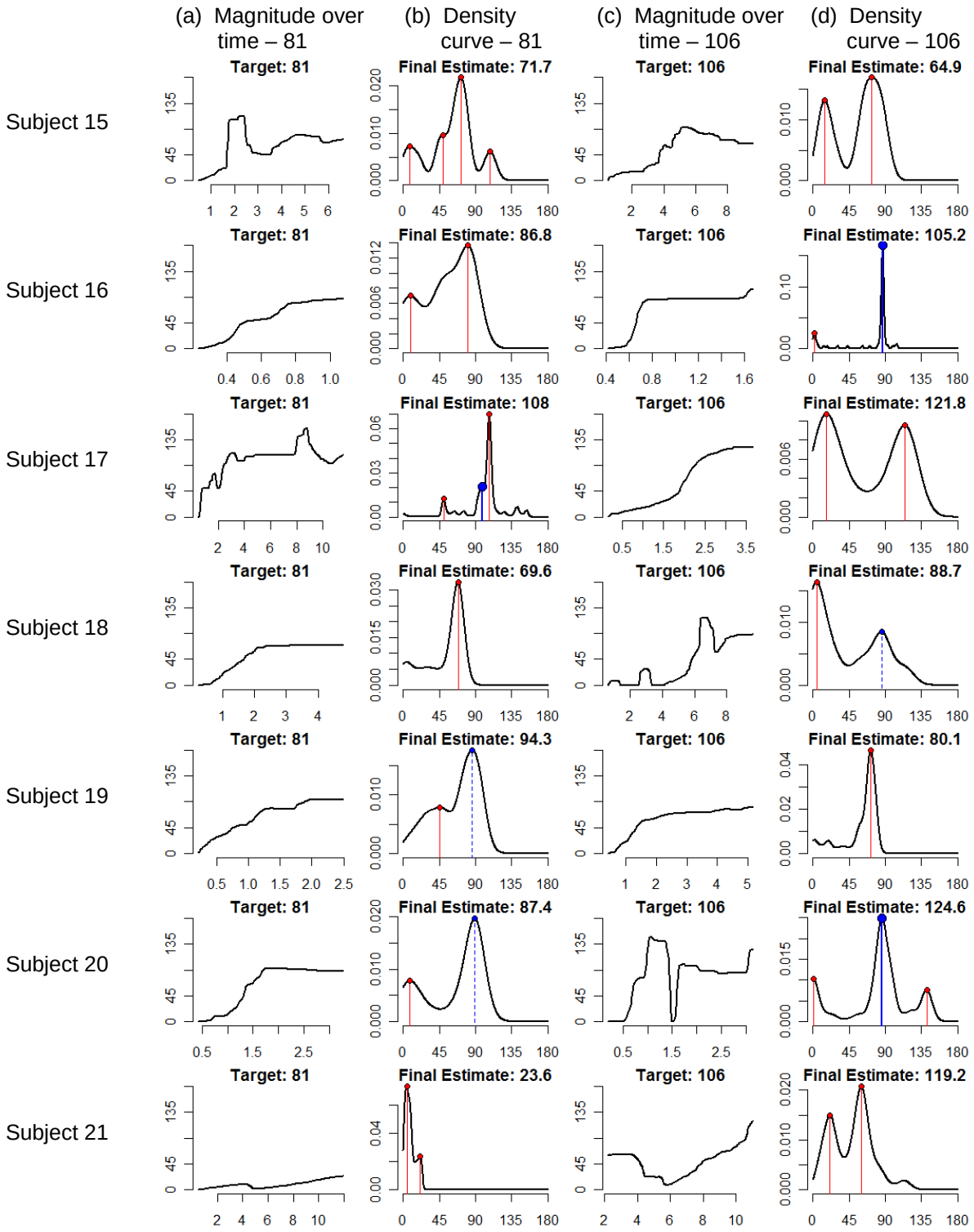
Trial curves and density plots (targets 81 and 106, subtest 1) – *Congruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 2.3

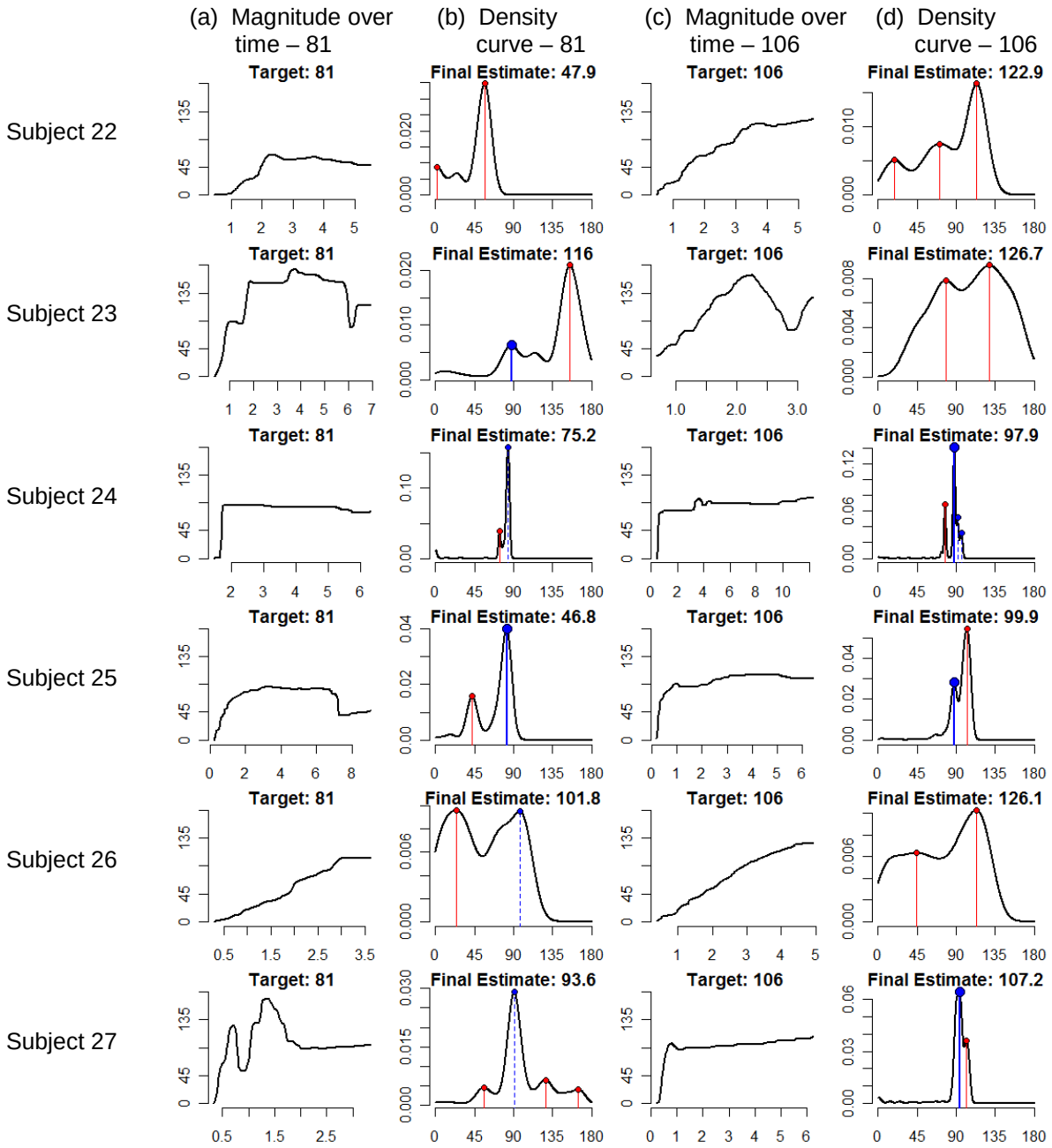
Trial curves and density plots (targets 81 and 106, subtest 1) – *Congruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 2.4

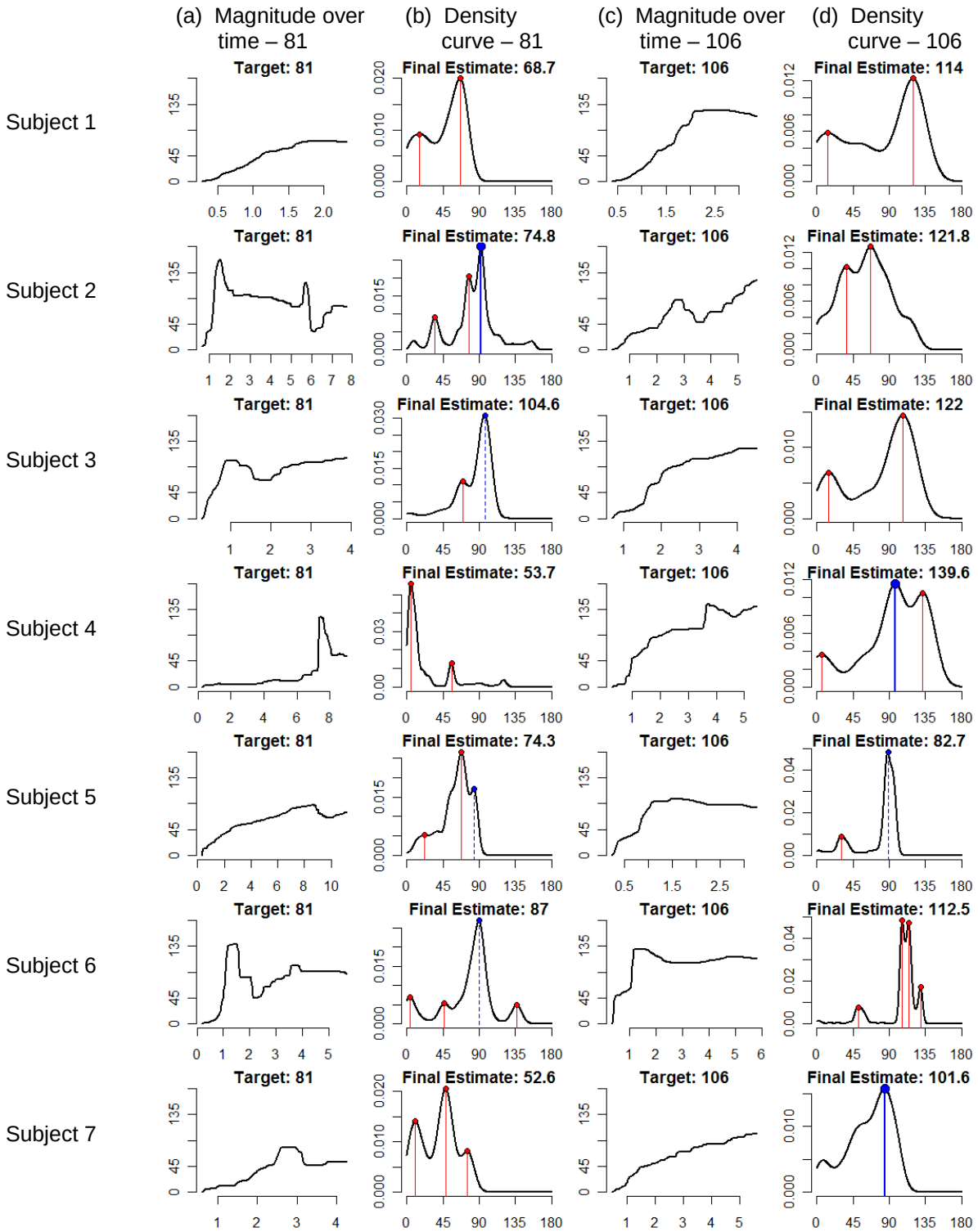
Trial curves and density plots (targets 81 and 106, subtest 1) – *Congruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 3.1

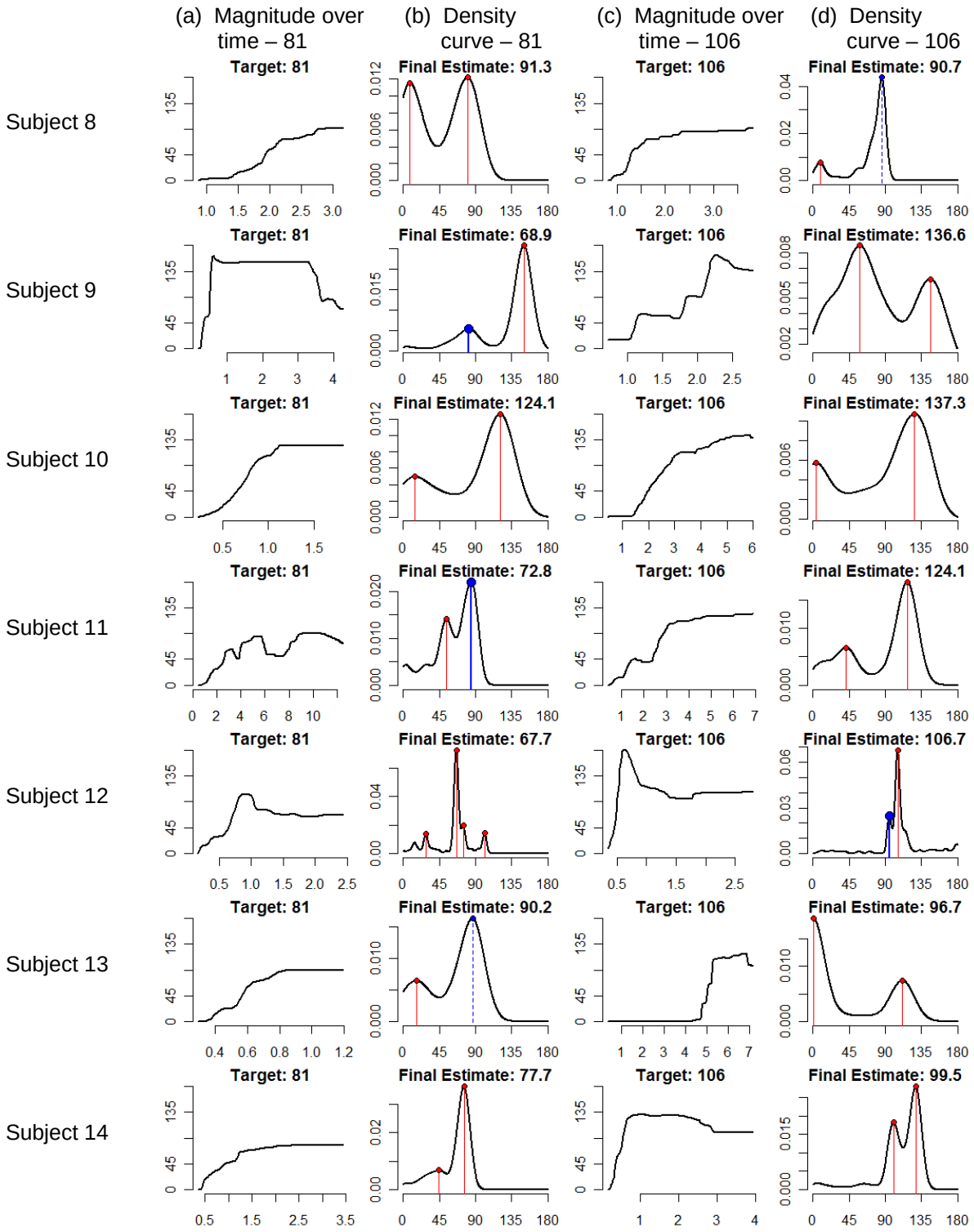
Trial curves and density plots (targets 81 and 106, subtest 1) – *Incongruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 3.2

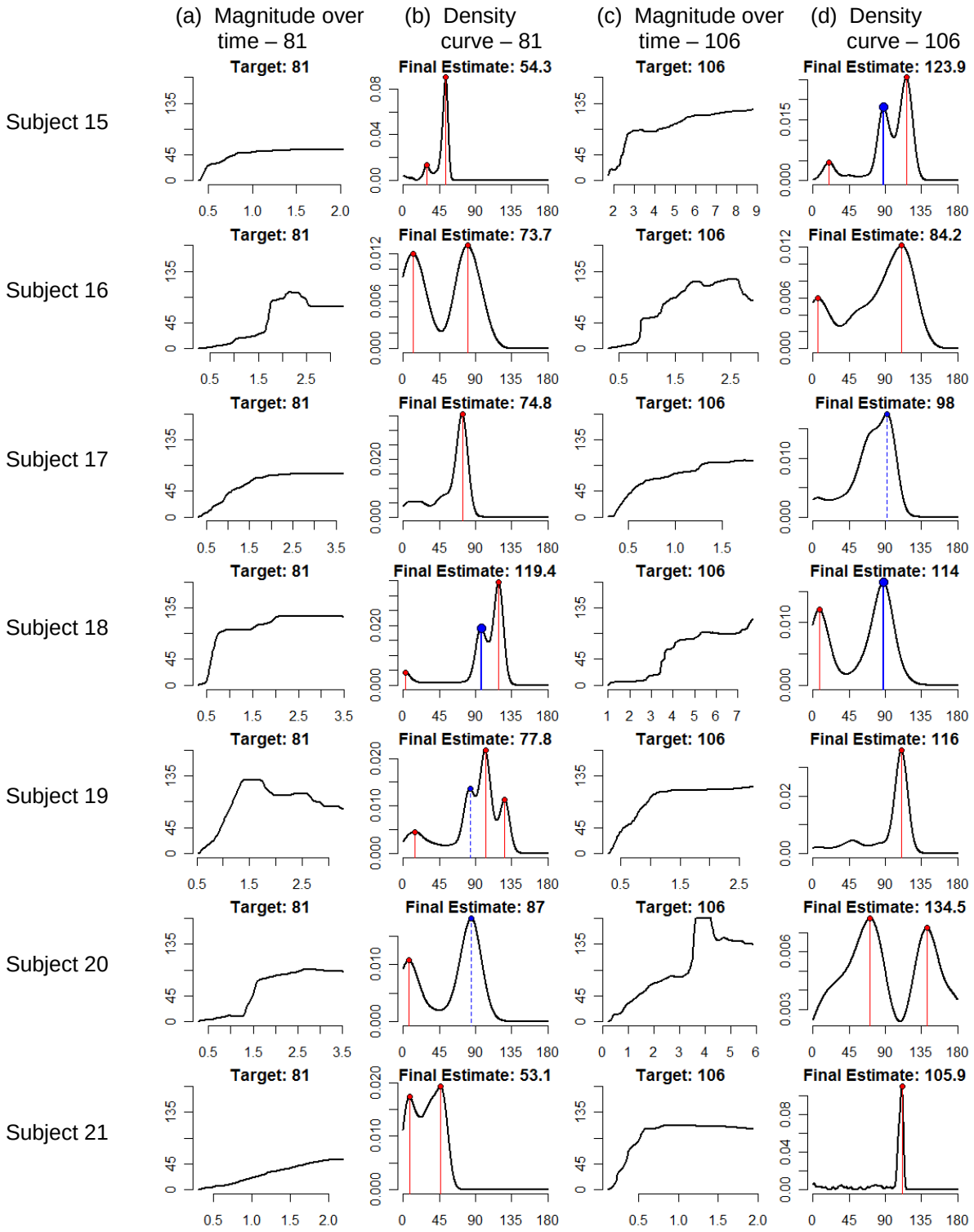
Trial curves and density plots (targets 81 and 106, subtest 1) – *Incongruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 3.3

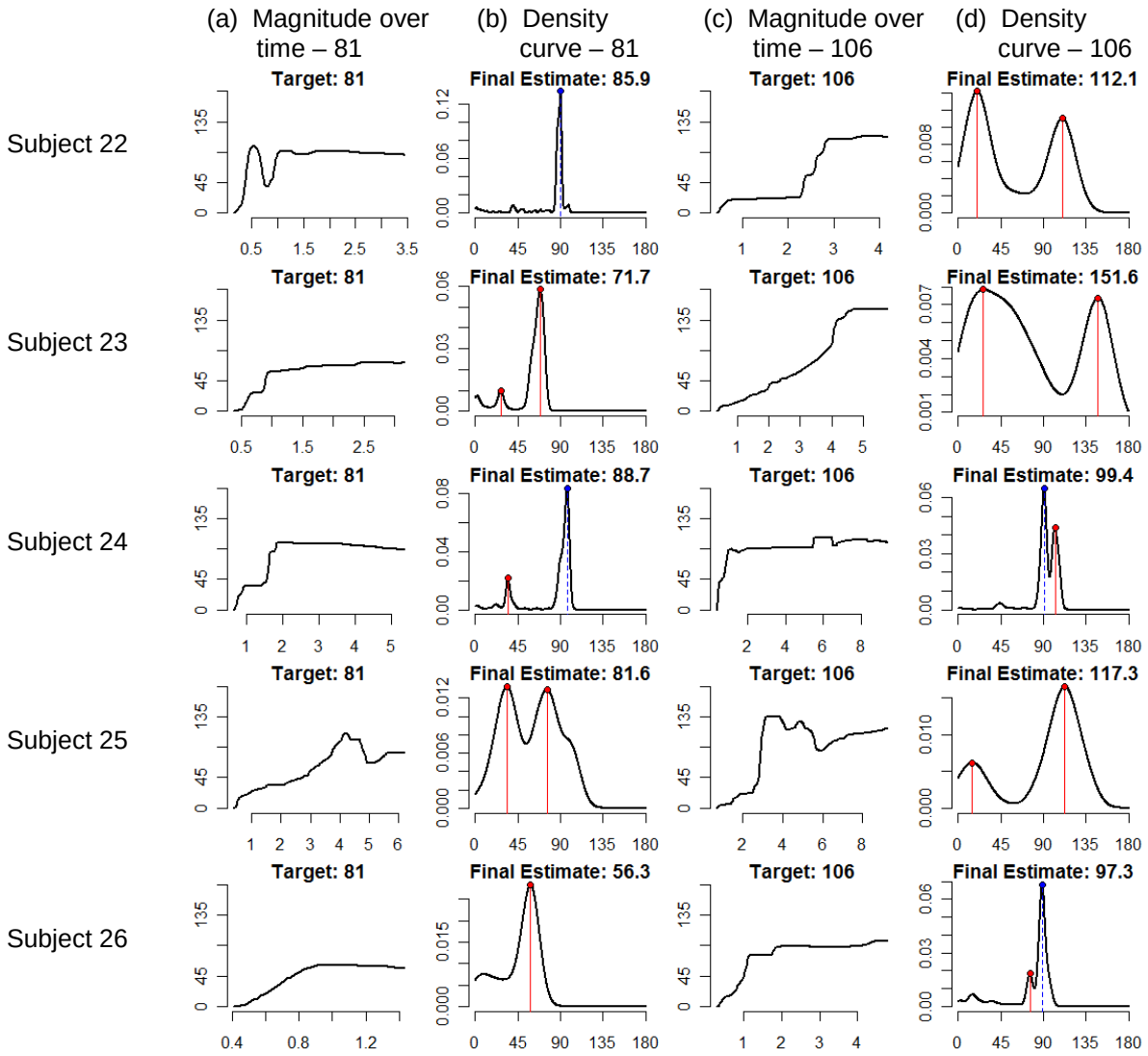
Trial curves and density plots (targets 81 and 106, subtest 1) – *Incongruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix H – 3.4

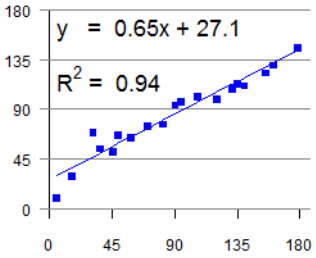
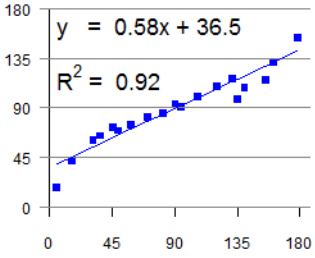
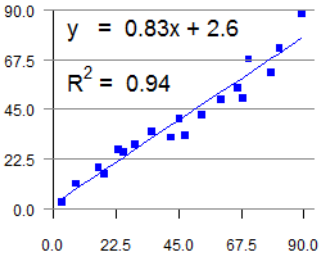
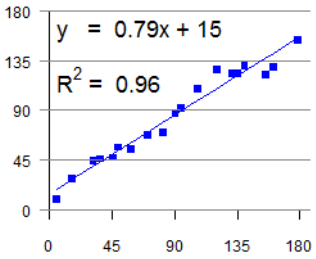
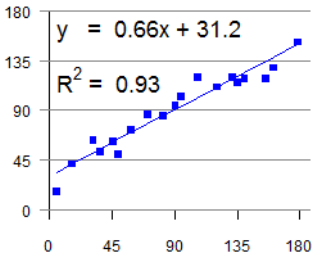
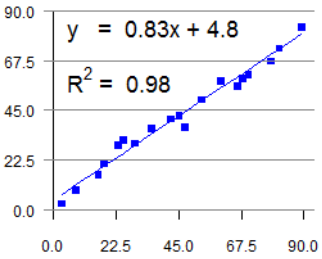
Trial curves and density plots (targets 81 and 106, subtest 1) – *Incongruent ruler* condition



For each subject, four graphs are shown: (a) In subtest 1, with trials targeting of 81, the current mouse cursor position (0 – 180) over time (in seconds). (b) A density plot of (a), where peaks represent (near) stopping points. Solid blue lines indicate a peak near 90. Dashed blue lines indicate a peak near 90 that is (suspiciously) close to the final estimate. Red lines indicate a stop outside of the margin of error from 90. (c) Same as (a), with trials targeting 106. (d) Same as (b), with trials targeting 106.

Appendix I

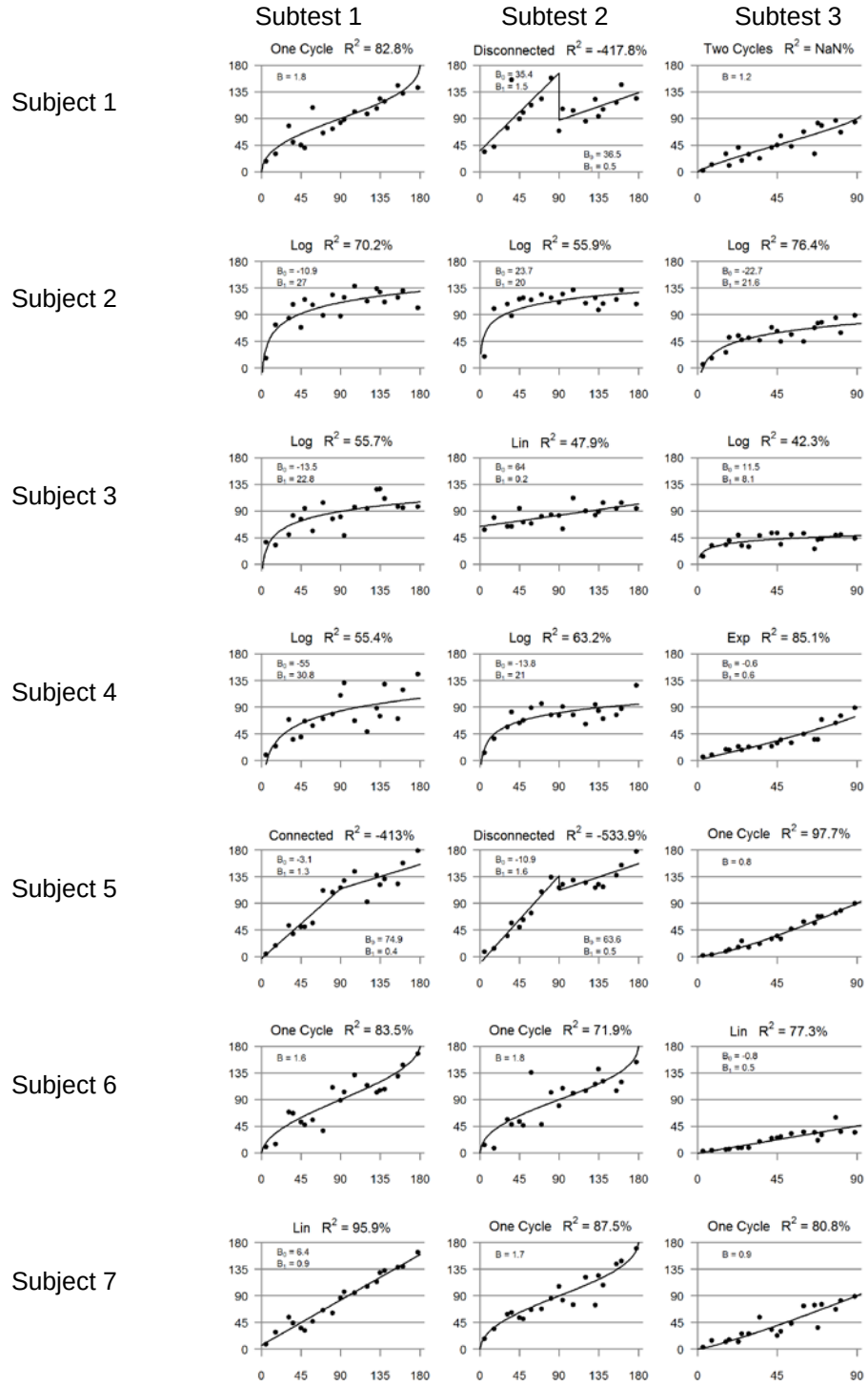
Median Estimates over Actual Magnitude – Experiment 2

	Equivalent subtest Full-length left-to-right 0 – 180	**Spatial transfer subtest** Half-length bottom-to-top 0 – 180	**Numerical transfer subtest** Full-length left-to-right 0 – 90
Congruent ruler			
Incongruent ruler			

Experiment 2 median estimates. X-axis represents actual magnitude, Y-axis represents median estimated magnitude. Columns display progression of subtests. Rows show different conditions.

Appendix J – 2.1

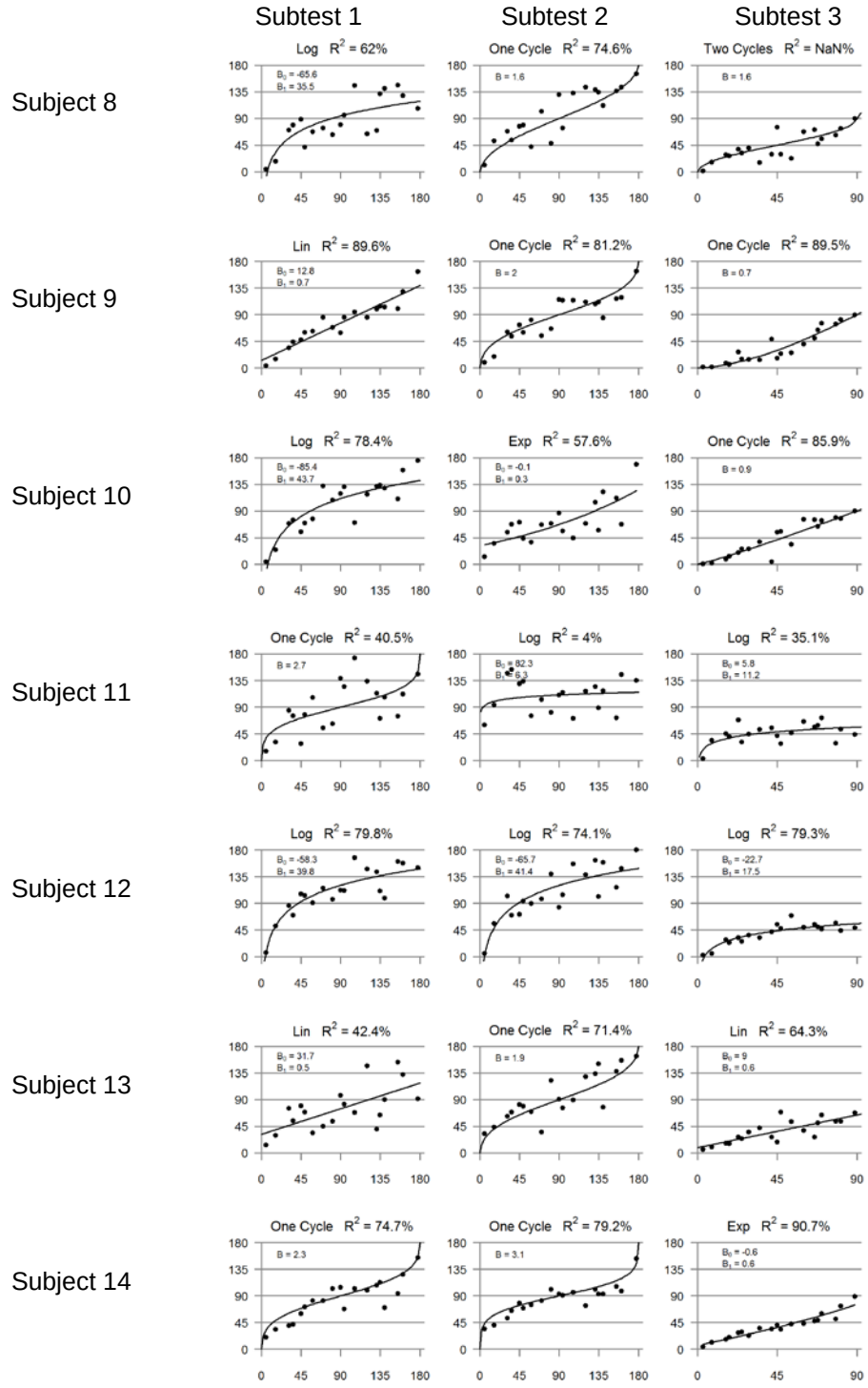
Individual Estimates over Actual Magnitude – *Congruent Ruler* Condition – Exp. 2



Description given on page 111.

Appendix J – 2.2

Individual Estimates over Actual Magnitude – *Congruent Ruler* Condition – Exp. 2

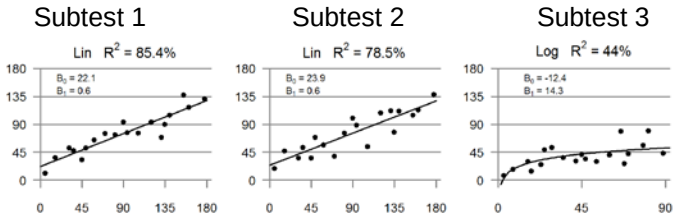


Description given on page 111.

Appendix J – 2.3

Individual Estimates over Actual Magnitude – *Congruent Ruler* Condition – Exp. 2

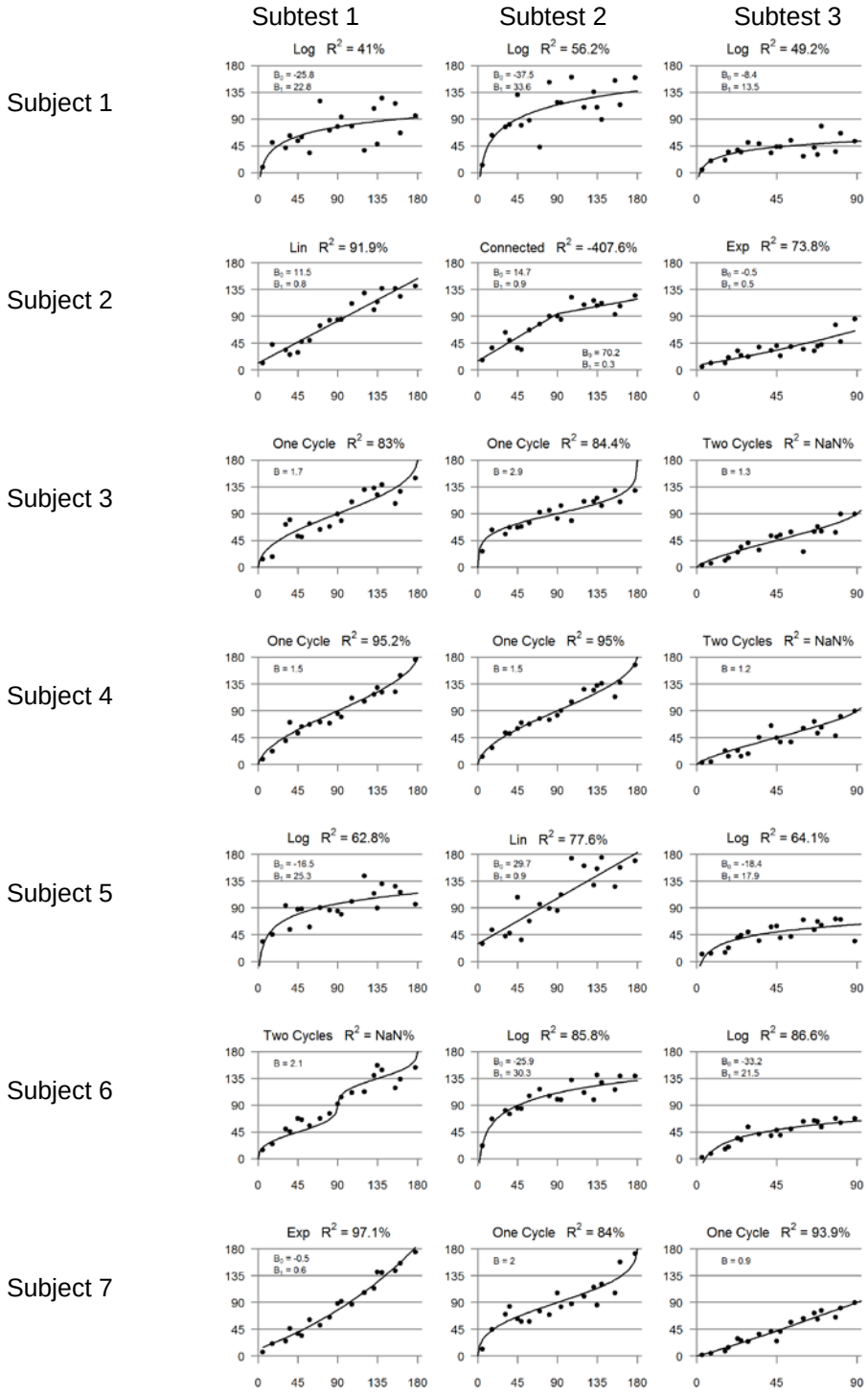
Subject 15



Description given on page 111.

Appendix J – 3.1

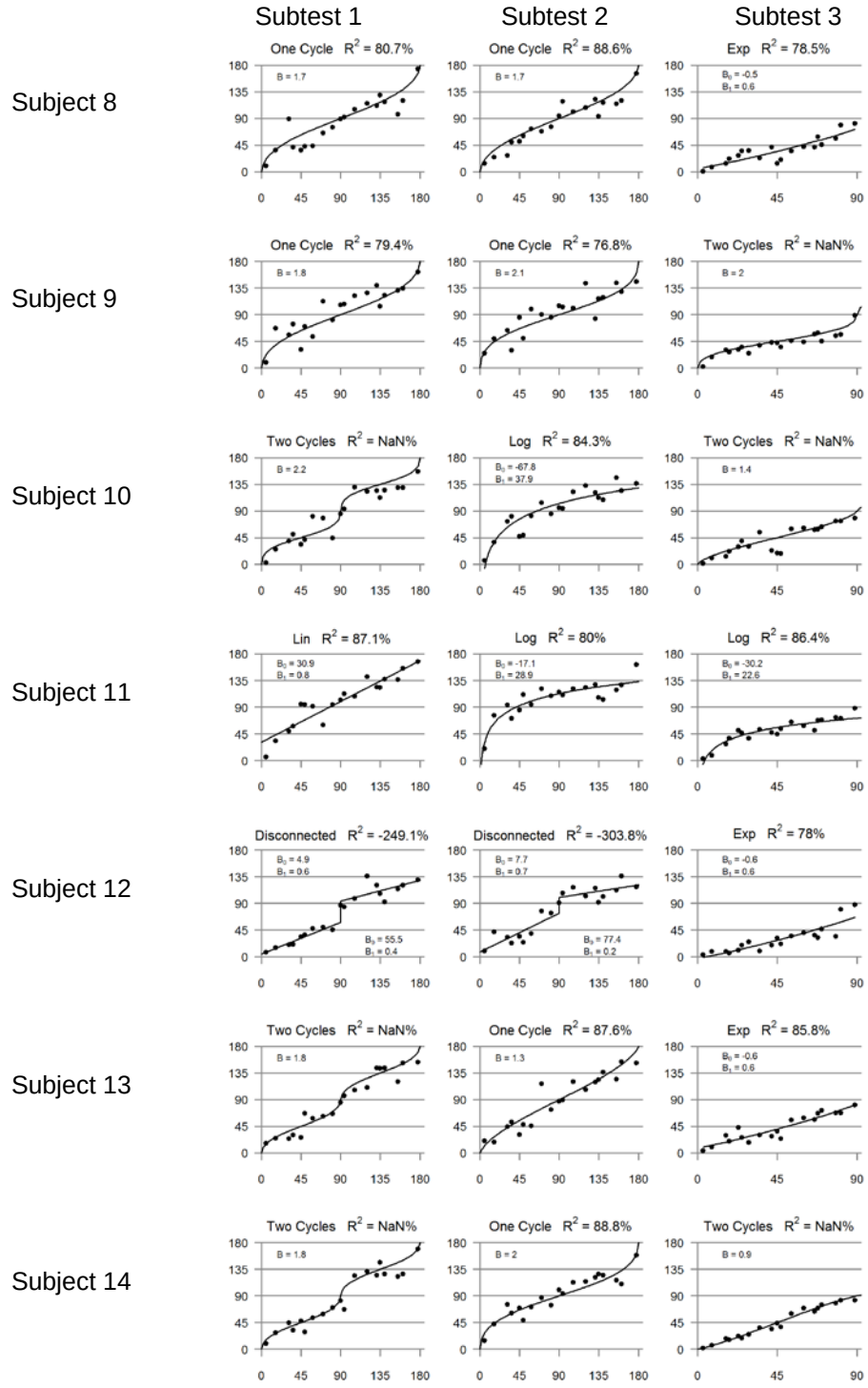
Individual Estimates over Actual Magnitude – *Incongruent Ruler* Condition – Exp. 2



Description given on page 111.

Appendix J – 3.2

Individual Estimates over Actual Magnitude – *Incongruent Ruler* Condition – Exp. 2

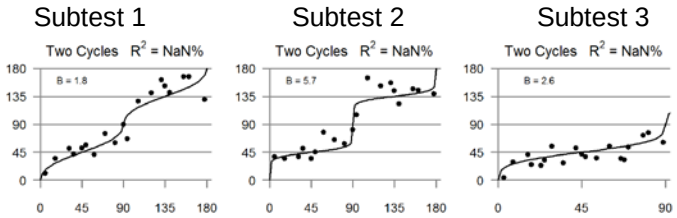


Description given on page 111.

Appendix J – 3.3

Individual Estimates over Actual Magnitude – *Incongruent Ruler* Condition – Exp. 2

Subject 15



Description given on page 111.

Appendix K – Experiment 2 Posttest Model Distributions

Frequency of linear and log models by condition and subtest

Subtest	Log		Linear	
	CR	*IR*	*CR*	*IR*
Equivalent	7	2	8	13
Spatial	5	4	10	11
Numerical	5	4	10	11

Frequency of linear, log, and power regression model types by condition and subtest

Subtest	Log		Linear		One Cycle		Two Cycle	
	CR	*IR*	*CR*	*IR*	*CR*	*IR*	*CR*	*IR*
Equivalent	6	2	4	4	5	4	0	5
Spatial	5	4	3	3	7	7	0	1
Numerical	5	4	3	3	5	2	2	6

Frequency of linear, log, and segmented regression model types by condition and subtest

Subtest	Log		Linear		Connected		Disconnected	
	CR	*IR*	*CR*	*IR*	*CR*	*IR*	*CR*	*IR*
Equivalent	7	2	7	12	1	0	0	1
Spatial	4	4	9	7	0	2	2	2
Numerical	5	4	10	11	0	0	0	0