2011

A Knowledge-Based Theory of Rising Scores on "Culture-Free" Tests

Mark C. Fox

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

A KNOWLEDGE-BASED THEORY OF RISING SCORES ON "CULTURE-FREE" TESTS

By

MARK C. FOX

A dissertation submitted to the
Department of Psychology
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2011

Mark C. Fox defended this dissertation on October 25, 2011.

The members of the supervisory committee were:

Neil Charness
Professor Directing Dissertation

Anne Barrett
University Representative

Colleen Kelley
Committee Member

Walter Boot
Committee Member

Carol Connor
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

ACKNOWLEDGEMENTS

I thank my committee members, Neil Charness, Colleen Kelley, Walter Boot, Anne Barrett, and Carol Connor, but I am most grateful to Neil for allowing me to confront the kinds of research questions that he must have known would consume more time, effort, and resources than I could have anticipated. I thank Anders Ericsson for helping me to appreciate that the reward of taking on big projects justifies the costly investment.

This project would not have been possible without the long-term dedication of motivated and talented research assistants, including, but not limited to, Alycia Marion, Leisa Leonard, Blake Cowing, Aaron Scheiner, and Christina Stenberg.

I am fortunate to have found myself surrounded by outstanding peers for the entire duration of my graduate career. Roy Roring, Edward Cokely, Carissa Zimmerman, Ryan Best, Celeste Doerr, and Ainsley Mitchum have been and remain superb colleagues and loyal friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Secular gains in intelligence test scores have perplexed researchers since they were documented by Flynn (1984, 1987), but few have attempted to understand them as a cognitive phenomenon. Gains are most pronounced on seemingly "culture-free" tests, which require analogical reasoning in the near-absence of familiar content, prompting Flynn (2007) to attribute rising scores to improvements in abstract reasoning conferred by a 20th-century emphasis on scientific thinking. Building upon Flynn's theory and Singley and Anderson's (1989) conceptualization of transfer as common productions, I propose that recent-born individuals have developed a relatively general procedural knowledge structure, or "weak method" (Singley & Anderson, 1989, p. 230), for analogical mapping. I test the theory first with archival data, and then with think-aloud verbal reports obtained while participants from two cohorts completed the Raven's Matrices, the test with the largest Flynn effect. Consistent with the theory, it is found that individuals from the earlier cohort are less able to map objects corresponding to higher levels of relational abstraction. Previous research suggests this weak method may be cultivated by learning to solve a wide variety of the kinds of unfamiliar problems that require an initial process of working through an example. The work identifies a plausible cognitive mechanism for the Flynn effect, makes testable predictions, reveals new insights into the cognition of matrix reasoning, and highlights the indispensible role of cognitive theories in advancing and testing cross-cultural generalizations.

CHAPTER ONE

THE ANALOGY OF "CULTURE-FREE" TESTS

Intelligence test scores rose dramatically during the 20th century (Flynn, 1984, 1987) and continue to rise in some parts of the world. Today's young adults in developed countries score about 15 points higher on full-scale IQ tests than their grandparents did at the same age, and as much as 25 points higher on some abstract reasoning tests (Flynn, 2007). Remarkably, it is the tests designed not to test knowledge that have seen the largest gains of all (Flynn, 1984; 1987). These "culture-free" tests consist primarily of abstract reasoning items that are typically devoid of any words, pictures, or other symbols that would be more recognizable to the average person today than they were a century ago. The Flynn effect has been described as "one of the most surprising, most intriguing—and potentially most important—findings in the recent psychology research literature" (Rodgers, 1998, pp. 337–338), and yet its cause remains uncertain after nearly three decades of research.

The search for prospective causes has focused on broad ecological changes, from improved education (Blair, Gamson, Thorne, Baker, 2005; Ceci, 1991), to better nutrition (Colom, Lluis-Font, & Andrés-Pueyo, 2005; Lynn, 1990; Martorell, 1998), to increasing availability of information (Barber, 2005), to shrinking sibship size (Sundet, Borren, Tambs, 2008). While each of these hypotheses is plausible, neither makes predictions that are clearly differentiable from the others in light of available data (Mingroni, 2007). Moreover, it is not readily apparent why any would be particularly conducive to gains on abstract culture-free tests as opposed to, say, general knowledge tests. Given that the Flynn effect is ultimately a cognitive phenomenon regardless of cause, identifying a plausible cognitive mechanism could help to isolate genuine ecological causes to the extent that the mechanism is more compatible with some alternatives than others. If one views the Flynn effect as a natural experiment, such an approach is a rare opportunity to gain new insights into the nature of problem solving and how it varies between persons and cultures.

The scientific context in which culture-free tests were conceived reveals some important clues about the demands they place on problem solvers. The associationist and early pioneer of intelligence testing, E. L. Thorndike, theorized that transfer of performance from one situation to another is limited to "identical elements" (Thorndike, 1922; Woodworth & Thorndike, 1901, p. 250). Thorndike did not clearly define "identical" (Singley & Anderson, 1989), but the most

straightforward interpretation of his claim is that elements are the superficial features of situations themselves. If this is true, removing familiar content (e.g., recognizable words, pictures of everyday objects) from test items would be expected to insulate a test from cultural bias. This assumption is significant in light of Charles Spearman's proposal that *eduction of relations* and *eduction of correlates* (Spearman, 1927, pp. 165–166), components of analogy, are the basis of human intelligence. Thorndike and Spearman's assumptions imply that the ability to solve analogies in the absence of familiar content is culture-free intelligence. In accordance with these assumptions, culture-free tests, although variable in appearance, have tended to share a common underlying structure of abstract analogy. Test-takers must solve analogies consisting of relations between objects that are not implied by the objects themselves. In a fascinating paper, L. S. Penrose, and Spearman's student, J. C. Raven (1936), described the items that would eventually comprise the Raven's Matrices (Ravens; Raven, 1938), the test with the largest Flynn effect:

> The fundaments can be spatially related figures; the eduction of the relations between the figures can then lead to the apprehension of a logical correlate. After this, the selection of the appropriate figure to complete the system is independent of technical ability. In tests of this form, all relevant training is entirely under the examiner's control. The relations to be educed can be made novel and problematic at every stage in the series: they need not depend on previous education (Penrose & Raven, 1936, pp. 97–98).

Most educated people can solve the analogy, *sun is to planet as nucleus is to_,* because they are familiar with the *analogs, solar system* and *atom.* Moreover, the *objects, sun*, *planet,* and *nucleus* have familiar *roles* or functions that are inherent to knowledge of suns, planets, and nuclei within the broader knowledge domains of solar systems and atoms (e.g., Salvucci & Anderson, 2001). To know a sun or a nucleus is to know that it attracts, and to know a planet is to know that it orbits because of the way that knowledge is organized (Collins & Quillian, 1969).

Abstract analogical reasoning items such as those on the Ravens are difficult in part because the *relations* between objects are not intrinsic to knowledge of the objects themselves. Many problem solvers would be stumped by the analogy, *&$B:B&$::T&T:$$_,* because no one's concept of *&* includes a role pertaining to *&*'s relation with *$* in this specific analogy. The principal distinguishing feature of this analogy is not its unfamiliarity per se, but the

indeterminacy of its roles and relations with respect to its objects. Indeterminacy characterizes any analogical or inductive inference (including those involving solar systems and atoms), but is mitigated substantially by the organization of declarative knowledge under ordinary circumstances. Analogies like this one are difficult because they withhold the information that is needed to take advantage of this organization. As Carpenter, Just, and Shell (1990) observed, the *rules* (presently defined as relations that are common to each analog) found on the Ravens are actually simpler than many encountered in context-rich domains such as political science. In fact, the rule in the problem above is the same as the most difficult rule found on the Ravens (Carpenter et al.'s distribution-of-two-values rule), but is no less simple or familiar than the principle used to sort one's socks.

While American and British psychologists were attempting to measure culture-free intelligence with content-reduced analogies, gestalt psychologists of the Austrian-German tradition such as Wertheimer, Koffka, and Katona were examining problem solving from a different perspective, and uncovering findings that challenged Thorndike's identical elements theory. Their work showed that problem solvers can often transfer a principle learned while solving one problem to superficially different problems with the same basic structure (Singley & Anderson, 1989). "To the gestalters, transfer occurred not through the piecemeal sharing of common elements but rather through the *transposition* of an entire structure" (Singley & Anderson, 1989, p. 9, original emphasis).

Adopting a representational framework descended from both behaviorism and gestalt psychology, Singley and Anderson (1989; Anderson, 1987[1]) concluded that transfer is the procedural knowledge common to multiple problems or tasks. This implies that a single procedural skill that is common to a variety of different tasks could potentially account for rising scores on multiple tests. The implications of Singley and Anderson's (1989) theory for understanding the Flynn effect are compelling in light of Flynn's (2007) claim that the 20th-century emphasis on scientific thinking forced 20th-century people to think more abstractly than their predecessors. A possible response to these demands was development of a generalizable series of procedural knowledge rules, or a "weak method" (Newell, 1973; Singley & Anderson,

[1] Anderson (1987) anticipates many of the ideas presented in Singley and Anderson (1989), but I refer primarily to the latter source simply because it is more comprehensive, presents additional data, and emphasizes transfer more explicitly.

1989, p. 230) of problem solving, that is more applicable to a problem's structure than its content. In this paper I propose that recent cohorts have developed a specific form of procedural knowledge for analogical mapping when relations between objects have no declarative association with the objects themselves, thus improving performance on abstract "culture-free" tests.

Before describing the theory in detail, I present a brief overview of prospective ecological causes of the Flynn effect and show that proliferation of some form of knowledge is more plausible and more consistent with available evidence than alternatives. Next, I argue that a change in procedural knowledge underlying mapping of objects in analogical reasoning could manifest as something very similar to the improved abstract reasoning described by Flynn (2007). I apply the theory to matrix reasoning tests, which show some of the largest Flynn effects, and present an analysis of data from over 3,000 young adults occupying three cohorts to test the prediction that mapping accounts for between-cohort variation in item-specific pass rates on a highly affected test. I briefly review relevant individual-differences findings before presenting a process tracing study to serve as a finer-grained test of the theory. Returning to the cognition of transfer, I propose a specific learning exercise that, according to prior research, could account for a substantial share of the gains in scores. I return to the question of ecological causes before discussing further predictions and relevant meta-theoretical issues.

What Caused The Flynn Effect?

Intelligence test scores rose dramatically in America and other developed countries during the twentieth century (Flynn, 1984, 1987) and continue to rise in other parts of the world (e.g., Brouwers, Van de Vijver, Van Hemert, 2009; Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Khaleefa, Abdelwahid, Abdulradi, & Lynn, 2008; Wicherts, Borsboom, & Dolan, 2009). The Flynn effect is most pronounced on tests that early- to mid-20th century researchers considered culture-free such as Cattell's Nonverbal Intelligence Test and the Ravens. Lynn, Hampson, and Millineux (1987) found the former to rise by nearly one-third of an IQ point per year between 1935 and 1985, and Flynn (1987) reported that the once-difficult Ravens, first published in 1938, was no longer normally distributed for Dutch military recruits by the early 1980s because of ceiling effects, forcing him to use medians as indicators of central tendency. Daley et al. (2003) reported that scores of rural Kenyans on this test rose by one-and-a-quarter to nearly two IQ points *per year* over the 14-year period between 1984 and 1998. Brouwers et al.'s

(2009) meta-analysis of Ravens scores over a 60-year period, based on 798 samples from 45 countries, showed that scores rose by about two IQ points per decade, yet the authors observed that "the Raven tests are still considered to be measures of intelligence that show less influence of confounding cultural factors on the cross-national differences than any other intelligence test" (p. 330). Culture in many countries has clearly changed during this time, and yet the tests purported to measure it (viz. crystallized intelligence) have seen relatively minor gains. How is it possible for scores to rise so quickly on the very tests that are *not* supposed to measure cultural changes?

Given the disproportionate effect sizes for abstract, culture-free tests, it is tempting to rule out otherwise plausible explanations such as learning, or even dismiss environmental hypotheses altogether. Some have suggested that nutrition played a major role (Lynn, 1990; Sigman & Whaley, 1998) as some evidence suggests that nutritional supplementation can raise test scores (Schoenthaler, Amos, Eysenck, Peritz, & Yudkin, 1991). However, the effect sizes of nutritional supplementation are relatively small (Flynn, 1999), and there is little regional or temporal correspondence between nutritional improvements and rising scores (Flynn, 1992, 1999b, 2007). Mingroni (2007) has suggested that the magnitude and stability of intelligence heritability estimates (Neisser et al., 1996)—heritabilities have stayed the same while scores have risen—imply a genetic cause, but Sundet, Eriksen, Borren, and Tambs (2010) observed a within-sibship Flynn effect for 69,000 Norwegian brother-pairs, which cannot be explained by a genetic change. Woodley's (2011) recent critique of Mingroni's (2007) theory provides additional evidence against a genetic account of rising scores.

Environmental explanations are not only compatible with stable heritabilities (Dickens & Flynn, 2001), but are *vastly* more plausible than biological alternatives. Genotypic changes, including the relatively fast variant of heterosis suggested by Mingroni (2007), are glacial compared to environmental stimuli that can dramatically affect a single individual within days or even hours. Within a typical range of between-subjects variation, ordinary learning confers improvements in performance that are orders of magnitude larger than any alternative, either biological or environmental. Merely taking the Ravens test can improves one's score by nearly one standard deviation on the same test as late as 45 days later (Bors & Vigneau, 2001), and inductive reasoning training can raise scores on similar tests rapidly and substantially (see Klauer & Phye, 2008). Assuming that the Flynn effect is not identical to typical practice or training

gains, the sheer magnitude of these effects implies, at the very least, that knowledge can raise scores if the proper environmental stimuli are available.

Knowledge is also consistent with Brouwers et al.'s (2009) meta-analysis of the Ravens, a test that has been administered to more people worldwide than any other except for the Wechsler Intelligence Scales for Children (WISC; Brouwers et al., 2009; Van de Vijver, 1997). Scores at any given time (i.e., when controlling for publication year) were found to be associated with educational age (years of education in the test sample) and educational permeation of country, both of which coincide with cultural factors such as economic development. Given that primarily young people have been tested (the mean age was about 17 years for the nearly quarter-of-a-million participants), often in only recently developing countries, it is more likely that these factors cause higher test scores than vice versa. A knowledge-based explanation is also compatible with evidence that variation in fluid intelligence test scores diminished, especially among the lower performing half of the sample distribution (Colom et al., 2005; Teasdale & Owen, 1989). Multigroup confirmatory factor analysis reveals violations of measurement invariance between cohorts (Must, te Nijenhuis, Must, & van Vianen, 2009; Wicherts et al., 2004) that are consistent with proliferation of one or more forms of knowledge.

The set of possible knowledge-based hypotheses is broad because it includes any hypothesis that assumes individuals from recent cohorts have learned something from their environments that helps them to score higher on tests. Thus, hypotheses as diverse as so-called test sophistication, improved education (Blair et al., 2005), and proliferation of video games (Greenfield, 1998) are all knowledge-based hypotheses. Yet, as distinct as they may appear, it is unclear that these hypotheses are even mutually exclusive given that no one has stated explicitly what test sophistication is or identified the mechanisms by which education or video games would raise scores on abstract reasoning tests. A more precise way of evaluating knowledge as a potential cause is to conceptualize it not as an ecological resource, but as a psychological phenomenon.

CHAPTER TWO

ABSTRACT REASONING AS PROCEDURAL KNOWLEDGE

Flynn (2007) surmised that everyday cognition in the modern world requires more abstraction than a century ago when agriculture and industry were the most common vocations, and the only symbols ordinary people dealt with were familiar letters and numbers:

> Before 1900 most Americans had few years of school and then worked long hours in factories, shops, or agriculture. Their world was largely concrete….Aside from basic arithmetic, non-verbal symbols were restricted to musical notation (for an elite) and playing cards (except for the religious)…After 1950…more and more people began to put on scientific spectacles. As use of logic and the hypothetical moved beyond the concrete, people developed habits of mind. They became practiced at solving problems with abstract or visual content and more innovative at administrative tasks….The scientific ethos provided the prerequisites for this advance (pp. 173-174).

Modern cultures emphasize abstract thinking more than cultures of yesteryear and cultures in the underdeveloped world (Flynn, 2007; Williams, 1998). Blair et al. (2005) and Baker et al. (2010) showed that mathematics curricula in particular have shifted from emphasis on rote repetition to inference-based learning, and have placed increasing demands on inductive reasoning since the first half of the twentieth century. I return to actual ecological causes later in the paper, but the more pressing issue is what it means to become better at abstract thinking. I propose that individuals from later cohorts have acquired a form of knowledge that enables them to map analogical objects without having to access declarative knowledge about these objects. This knowledge is acquired from multiple learning contexts that emphasize deep, functional, structural relationships. The theory is an application of a broad theory of transfer proposed by Singley and Anderson (1989), and envisages cohort differences in analogical reasoning as changes in procedural knowledge. This "know how" knowledge can be formalized as rules called *productions* that dictate which processes are executed as a function of an active goal state in working memory (Anderson, 2007). The proposal requires, but deemphasizes the role of, declarative or "know that" knowledge, which is more akin to the everyday conception of knowledge characterizing more obvious explanations of the Flynn effect that have received little

empirical support. The consolidation of declarative knowledge into many "chunks" plays a key role in the development of domain-specific representations in many domains such as physics (Chi, Feltovich, & Glaser, 1981), but occupies a secondary role in the present theory. Instead, problem representation is regarded primarily as an emergent property of relatively general productions for analogical mapping of objects.

I begin with a basic overview of procedural knowledge and how it relates to analogy, intentionally leaving crucial questions about transfer unanswered until the general discussion where they can be considered from a more informed frame of reference. Because the theory must incorporate findings from several areas of cognition, I refer to the well-known and integrative Adaptive Control of Thought-Rational (ACT-R; Anderson, 2007) theory to substantiate abstract concepts and processes. This serves more than a purely communicative purpose because ACT-R is the foundation for both Singley and Anderson's (1989) theory of transfer as well as a theory of analogical reasoning that lends itself to specific applications such as the present one (Salvucci & Anderson, 2001). The latter is simultaneously versatile enough to be adapted to culture-free tests, and sufficiently compatible with other theories of analogical reasoning (e.g., Doumas, Hummel, & Sandhofer, 2008; Hummel & Holyoak, 1997) to obviate the need for a new alternative.

A Procedural View of Skill Acquisition

Conceptualizing test performance in terms of skill acquisition eliminates some of the theoretical constraints imposed by the interpretation of intelligence tests as measurement instruments, which is potentially misleading, especially in the context of the Flynn effect. There is clearly a tendency to interpret the actual variation observed in one population as representing a test's full range of *potential* variation. This is problematic because, just as every educated adult knows how to read in modern developed countries, there is a conceivable population in which little variation in scores is observed for tests that reveal substantial variation in samples of early-21st-century persons in developed countries. I adopt a perspective compatible with Borsboom, Mellenbergh, & van Heerden's (2004) conception of validity whereby between-subjects variation in a population is considered incidental to the means by which any one individual achieves an absolute level of performance.

Cognitive psychologists often distinguish between declarative knowledge, or knowledge of facts, and procedural knowledge, or knowledge that guides execution of tasks (Chi & Ohlsson,

2005). Procedural knowledge can be conceptualized as production rules (productions) or if-then statements that guide the execution of one or more processes in response to a goal and other information that is active in working memory at a given time (Anderson, 2007). Declarative knowledge plays an important role in the early stages of skill acquisition, as evidenced by young children who must enunciate the individual phonemes within words while learning to read, and music students who can identify the notes that correspond to every line and space of a staff, but cannot translate this notation into motor movements quickly or accurately enough to produce coherent music. From a skill acquisition perspective, proficiency at skills such as reading is the result of procedural knowledge adapting so as to optimize the efficiency and accuracy of performance. Productions, the "elements" of procedural knowledge (Singley & Anderson, 1989, p. 138), figure in at the outset of skill acquisition to initiate processes, but at this stage are too numerous, weak, and inconsistent to guide efficient execution of any one task (Anderson, 1987; Newell, 1973). With practice, a more optimal list of productions is compiled through elimination of single productions, consolidation of multiple productions, and greater activation of productions that are most relevant. For example, production compilation in the ACT-R architecture consolidates productions that fire successively into single productions (Anderson, 2007; Taatgen & Lee, 2003). Eventually, unnecessary processes are minimized or even eliminated entirely, resulting in faster performance with fewer opportunities for error (Taatgen & Lee, 2003).

Adaptation of procedural knowledge can occur long before expert levels of proficiency are reached even in the absence of new declarative knowledge. Numerous studies show that participants can gradually learn to adapt to rule-based algorithms without acquiring explicit declarative knowledge of the algorithm (Anderson & Fincham, 1994; Berry & Broadbent, 1984, Willingham, Nissen, & Bullemer, 1989). For example, Berry and Broadbent (1984) placed participants in hypothetical rule-based environments where they managed either a sugar factory or an interpersonal interaction. Both environments were governed by a mathematical function that was not revealed explicitly to participants. Performance improved rapidly in both scenarios even though participants could not report the rule, suggesting that improvements were not dependent on acquisition of long-term declarative knowledge. An even more dramatic example (Anderson & Fincham, 1994) is the well-established finding that amnesic patients are able to learn rule-based skills such as a mathematical algorithm, applying it accurately in the absence of

awareness that they possess any new knowledge (e.g., Charness, Milberg, & Alexander, 1988). From a skill acquisition perspective, these examples are, at least in theory, the result of a compilation process that optimizes procedural knowledge through modification of productions (Anderson & Fincham, 1994).

The claim that productions can become optimized to a specific task or skill is relatively straightforward, but some have suggested that general problem solving heuristics such as hill-climbing and means-ends analysis are a realization of the same compilation process (Anderson, 1987, Newell, 1973). Singley and Anderson (1989) note that transfer of broadly-applicable procedures such as these would be difficult to observe with the prototypical transfer methodology because they are too overlearned to be improved substantially. In light of this consideration, a rejection of the null hypothesis in a typical transfer study is capable of providing strong evidence of transfer, but failure to reject does not necessarily imply that transfer is impossible, implausible, or even uncommon.

Singley and Anderson's (1989) theory assumes procedural knowledge from one task can be transferred to other tasks with a similar procedural structure, that is, tasks that require the same productions. In fact, they suggest that transfer is synonymous with the productions common to both tasks. By their account, productions "can serve as the basis for substantial transfer even when a superficial analysis of similarity yields no common elements" (p. 138). The authors tested the theory in several domains, finding that participants experienced substantial savings when learning a new text editing program following training on another that was expected to require many of the same productions even though the two programs were distinct in their superficial content such as the actual commands entered in to the computer. Within the domain of calculus, participants taught integration transferred the operators they had learned to differentiation. In both cases process-relevant data suggested that the same productions guided performance in both the training and transfer tasks.

Almost by necessity, procedural knowledge that is applicable to multiple tasks cannot be sufficient by itself to optimize performance on any one task in particular, which is why Singley and Anderson (1989, p. 230) described such knowledge as a "weak" method of problem solving. Transfer can only be expected to occur between tasks that place similar constraints on execution and performance. The analogical reasoning underlying performance on culture-free tests is, in principle, an ideal candidate for a weak method because it relies on a basic conceptual structure

that is common to many different types of problems (Salvucci & Anderson, 2001), the vast majority of which are not themselves formal analogies. Analogy plays a foundational role in human cognition (Penn & Holyoak, 2008), but before considering how procedural knowledge for analogy is acquired, it is first necessary to consider the process of analogy itself in some detail, including its relationship to the task-demands of culture-free tests.

Analogical Mapping

As stated, an analogy consists of two or more analogs that are similar to one another in one or more ways. These analogs are composed of objects, defined as the immediately apparent properties of an analog, or the "pieces" that constitute an initial representation of the analog. Relations are the relationships between objects within an analog. Mapping is the process of associating objects in one analog to corresponding objects in another based on a common role served by an object in both the source and target analog. The term, *rule,* is used frequently in the matrix reasoning literature that I review below. The terminology I use here can be reconciled with that usage by designating a rule as a relation that is common to both analogs. In the analogy, *BAC:DEF::213:456, BAC:DEF,* and *213:456* are analogs, numbers and letters are objects (to literate and numerate people), and *order* is an example of both a relation and rule that is presented, for the sake of simplicity, without the specific arguments that would refer to the roles, *2, 1, 3, 4, 5, 6*. The analogy, *ELECTRON:NUCLEUS::PLANET:SUN,* offers a coherent representation of objects at the level of whole words instead of letters. The rule, object *attracts orbiting* object, applies to the roles, *planet, sun, electron,* and *nucleus.*

There are two primary ways for analogs to be similar: either their objects are immediately similar to one another, or their objects are immediately dissimilar to one another. In both cases roles are similar to one another, but in the latter case, *only* roles are similar; that is, objects in both analogs do not seem similar until one considers that they "behave" the same way or "perform" the same function (Gentner, 1983; Penn, Holyoak, & Povinelli, 2008). For present purposes, I commit to the terms *similar* and *dissimilar* to distinguish between analogies that are similar with respect to initial representation of objects, and those that are superficially dissimilar with respect to initial representation of objects. In order to avoid confusion, it is important to remember that the similarity of objects is determined not by any objective or external criteria (such criteria do not exist), but by a problem solver's initial representation, which, depending on the analogy, may or may not be the same as another problem solver's. Although it may appear

that I am inviting confusion by describing objects that are functionally similar as *dissimilar* simply because their similarity is not readily apparent, this language makes it possible to commit to the object as the privileged unit of analysis while acknowledging that higher-level representations are needed for mapping. This is necessary because the problem solver, not the problem itself, is the focus of this paper.

Most theories of analogy (e.g., Gentner, 1983; Hummel & Holyoak, 1997; Salvucci & Anderson, 2001) assume that mapping depends on the organization of declarative knowledge about analogs and relations between objects. Source and target domains (relevant declarative knowledge domains) of familiar analogs such as solar systems and atoms have the structure of semantic networks (Collins & Loftus, 1975; Collins & Quillian, 1969), which facilitates retrieval of common relations between objects (Hummel & Holyoak, 1997; Salvucci & Anderson, 2001). To know a planet or an electron is to also know that these objects orbit, and to know a sun or a nucleus is to also know that these objects are orbited. The former and latter objects can be mapped, in part, because their roles are intrinsic to their existence as concepts (note throughout that a concept *does not exist* except as a cognitive entity).

The present theory disputes none of this, but distinguishes between ordinary analogies and the *self-contained analogies* found on culture-free tests. Self-contained analogies are, by design, artificial in the sense that relations between objects are not associated with the objects themselves based on the organization of declarative knowledge.[2] Self-contained analogies consist of objects that may or may not be familiar (symbols, novel shapes, etc.), but crucially, relations between these objects are not implied by the objects themselves. Consider the analogy, *&$B:B&$::T&T:$$_*. No one possesses declarative knowledge of a relation between ampersands and dollar signs that is relevant to solving this analogy. Relations in self-contained analogies *do not exist* to a problem solver, and therefore must be induced from the properties of the analogy itself, which requires the development of a specific skill.

In what follows, I refer to Salvucci and Anderson's (2001) versatile path-mapping theory of analogy, which operates within the ACT-R architecture, and is adaptable to a variety of problem domains. The path-mapping theory is compatible with firmly established findings, and is comparable in function to Hummel and Holyoak's (1997) influential Learning and Inference

[2] This distinction is merely a conceptual heuristic. Analogies cannot be self-contained in any absolute sense that is independent of declarative knowledge.

with Schemas and Analogies (LISA) model and Doumas et al.'s (2008) more recent successor to LISA that I describe briefly in the general discussion. The basic path-mapping mechanism can be adapted to any type of analogy by means of higher-level procedural knowledge or *organizational knowledge* that governs the application of path-mapping in accordance with instructions and constraints of specific tasks (Salvucci & Anderson, 2001). What follows is more accurately conceptualized as a claim about this organizational knowledge than mapping itself. For this reason, I do not introduce additional terminology that is specific to path-mapping (see Salvucci & Anderson, 2001, pp. 76-77).

To preview, it is assumed that mapping in self-contained analogies is driven primarily by superficial similarities (initial representations of objects), and that abstract relations must be induced actively. Mapping of similar objects can be accomplished much like mapping in ordinary analogies except that the source and target analogs themselves must serve as the source and target domains (the relevant declarative knowledge) for identifying roles and relations. This makes it possible to achieve mapping by simply representing objects as their own roles and analogs as their own relations. However, this approach is limited to superficial analogies, and cannot support mapping of dissimilar objects. Mapping of dissimilar objects requires an active representation of roles as "unknowns" so that more abstract relations can be generated and/or retrieved.

In path-mapping (Salvucci & Anderson, 2001), objects are assigned roles that correspond to their functions within a source domain. A role is an unknown or open slot within the path-mapping productions that is capable of referring to any one of multiple functions of an object depending on relations between objects in a source analog (*sun is to planet*) and target analog (*as nucleus is to_*). Relations are determined by the organization of knowledge in source and target domains such as solar systems and atoms. In the case of *sun is to planet as nucleus is to__*, roles for the objects, *sun* and *planet,* are selected tentatively within the source domain (solar systems) by mapping analogous paths to the roles corresponding to the object's most general relation (*cause and effect*). Roles of more specific relations such as *attracts and orbits* are mapped along the path. Next, these tentative roles in the source analog are mapped to the most similar and specific roles in the target analog that are common to both source and target analogs. The higher-level (less specific) roles such as *cause* and *effect* are abandoned. Thus, *sun* is mapped to *nucleus* because these objects share the role, *attract,* and *planet* is mapped to *electron* because these

objects share the role, *orbit.* Salvucci and Anderson (2001) show that this same basic mechanism captures the incremental approach of human problem solvers, and is flexible enough to accommodate a wide range of analogical reasoning goals.

Mapping Similar Objects

In the preceding analogy, the objects, *sun* and *nucleus,* share common roles such as *attracts* and *center*. However, the object, *sun,* could conceivably occupy a role as specific as *sun* (sic) in a superficial analogy that compares two solar systems. The implication is that path-mapping would tend to assign objects to serve as their own roles when the source domain and target domains are the same. This is important because it implies that path-mapping requires little additional organizational knowledge (much of which would pertain to task instructions) to accomplish mapping in any self-contained analogy in which the roles of objects are identical to their initial representation as objects. Phrased another way, the basic path-mapping productions are sufficient (or nearly sufficient) for mapping objects in a superficial analogy because the source and target analogs are, *by definition,* the same as the source and target domains when objects are synonymous with their roles. There is no need to retrieve relevant declarative knowledge about roles and relations, but only because the representation of objects happens to correspond perfectly with the roles of these objects.

Consider the analogy *&B:#E::&B:#_*. The "relations" are synonymous with the analogs, *&B:#E,* because the roles are synonymous with the objects, *&, B, #,* and *E.* The rule (which is the same as the relations) can be inferred precisely because objects are synonymous with their roles. Note, however, that the rule does not generalize beyond this analogy, and for the very same reason: objects are synonymous with their roles. This synonymy between object and role is the precondition for analogs to be synonymous with their knowledge domains, thus obviating the need to retrieve non-existent declarative knowledge about relations between objects such as ampersands and letters. Although the rule, *&B:#E,* may seem contrived, objects that serve as their own roles are the norm for rules comprising easier items on tests such as the Ravens.

Mapping Dissimilar Objects

Because superficial mapping entails assigning objects to be their own roles, the procedure is inflexible in the same way that a numerical constant is inflexible when attempting to solve a geometry or physics problem that demands a formula containing variables such as the

Pythagorean theorem or Boyle's law. Such a procedure will only work in those rare cases when the constants happen to correspond to the actual values within the problem at hand. The fundamental distinction between analogies and geometry problems (aside from the fact that analogy subsumes all such problems as a more general category) is that an analogical procedure that assigns objects as their own roles when objects are dissimilar results not only in failure to solve the problem, but more catastrophically, failure to achieve even a coherent representation of *what the problem is.* This is because mapping is not just a sub-goal of analogy but an essential determinant of the problem's structure (this makes more sense when considering that formal analogical notation is usually not available as a heuristic for establishing structure). Unlike the content of other kinds of problems that rely on deductive logical principles (e.g., addition and subtraction), *the content of an analogy cannot be divorced from the means by which the analogy is solved.* [3] In other words, there is no algorithmic procedure for solving an analogy that does not presuppose the correctness of the problem solver's representation of roles. This indeterminacy is not a major problem when solving ordinary analogies because initial representation is ideal for eliciting retrieval of appropriate roles and relations. As the solar system and atom example shows, spreading activation is a powerful bootstrapping mechanism that reduces the number of possible roles to a few good candidates. In contrast, indeterminacy poses a significant problem when solving self-contained analogies with dissimilar objects because there is no source or target domain. A corollary of this absence of domain knowledge, as I suggest below, is that instructions and practice items alone cannot be expected to prepare problem solvers for a test of self-contained analogies.

A more flexible approach is needed to identify roles and relations when objects are dissimilar, one that allows role to vary like an unknown value. Although the basic path-mapping productions do treat roles as unknowns to be retrieved from the source and target domains, roles cannot be retrieved if the source and target domains do not exist. However, there is an alarmingly simple solution: additional bootstrapping (above and beyond ordinary spreading activation) in the form of productions that, functionally speaking, "acknowledge" that roles and relations are unknowns by retrieving and testing prospective roles and relations that defy initial representation

[3] This is a reference to the problem of induction, which implies that inductions cannot be justified on purely logical grounds (particulars cannot imply universals because the membership criteria of categories are indeterminate), although the decision to generalize may often be defended on pragmatic grounds.

of objects. In ordinary language, this knowledge corresponds to actively searching for coherent roles, but more subtly, "understanding" that *roles and relations are not necessarily consistent with initial representation.*[4] Although roles and relations are not immediately apparent, the analogy contains enough information to stimulate retrieval of the simplest and most generalizable among previously acquired roles and relations using ordinary retrieval or pattern recognition mechanisms. For example, a very common relation such as *number of x* in *&$B:B&$::T&T:$$_* may be recognized, or retrieved automatically even if this relation is not sufficient by itself for mapping the objects. Immediately apparent roles and relations can be altered or combined if they do not allow for mapping of relevant objects. Note that testing need not be considered an additional step, but rather, a process that occurs automatically when attempting to map objects in accordance with a prospective role (see footnote 4). This trial-and-error or trial-and-modification process makes it possible to establish roles that can be used to map the remaining objects in much the same way as objects would be mapped in ordinary analogies. What distinguishes this procedure from the superficial procedure above is that it allows roles to remain unknowns until some criterion (successful mapping) has been met.

The following productions, presented at a grain-size compatible with the present discussion, capture the process of generating and testing roles and relations by an architecture that acknowledges the indeterminacy of roles and relations.

1. IF the goal is to map objects, AND a prospective role is unavailable, THEN identify a prospective role (or relation) that has not been used or rejected.
2. IF the goal is to map objects, AND a prospective role is available, THEN map the other relevant objects according to their roles within the applicable relation.
3. IF the goal is to map objects, AND a prospective role is available, AND mapping was executed successfully, THEN accept the new target object.
4. IF the goal is to map objects, AND a prospective role is available, AND mapping was not executed successfully, THEN reject the role and its corresponding relation.

[4] It is reasonable to interpret this as a metacognitive consideration, but this interpretation is not necessary in principle and is likely to mischaracterize skilled problem solvers who have proceduralized the process.

When applied to the superficial analogy, *&B:#E::&B:#_,* the only role is the objects themselves, *&, B, #, and E,* which correspond to the rule, *&B:#E.* Applying production 2 maps objects as two versions of the same shape. In a more abstract analogy, *&$B:B&$::T&T:$$_,* no relation is immediately apparent except, perhaps, *number of x,* which can be applied successfully within analogs, but does not apply to *T* and *B* as a rule. A modification may reveal that the relation, *two of a kind,* can apply to every object in both analogs. This relation and rule is necessarily more abstract because its role, *pair,* subsumes the more concrete roles *&, $, B, and T,* and returns that the missing object is *&.* The most parsimonious roles and relations account for the most objects with the fewest relations. I elaborate more on the distinction between concrete and abstract roles and relations below by applying the same basic mechanism to matrix reasoning. The emergent property of procedural knowledge that allows roles to remain unknown is a higher-level analogical representation.

Application to the Flynn Effect

It is now appropriate to briefly consider the mechanism in relation to the Flynn effect. I am proposing that the Flynn effect was caused by proliferation of an analogical skill that is realized via compilation of productions, but I have not yet considered how these productions are acquired. According to this proposal, initial gains in test scores, appearing some 80 years ago (Flynn, 1984, 1987), may be the result of more individuals mastering the basic mapping skills captured by Salvucci and Anderson's (2001) path-mapping theory.[5] Over time, a greater proportion of individuals acquired procedural knowledge needed to generate and test possible roles and relations effectively. In ordinary language, such knowledge includes, for example, functional knowledge that one is not *supposed* to be able to simply retrieve relevant information from memory. One implication of this is that comprehension of instructions does not imply possession of a generalizable task representation because participants cannot know ahead of time what constitutes a role. Obviously, the same claim can be applied to any task with instructions and practice items because both are always analogous to the problems that follow. However, this

[5] I assume that theories of analogy tend to overestimate the analogical skills of the average adult when average is defined broadly enough to include adults like those mentioned by Flynn (2007) in the passage above. Test scores of individuals born prior to 1920 suggest that many performed worse than would be predicted by assuming they treated objects as their own roles. However, it is difficult to ascertain whether this is the result of an analogical deficiency or an analogy-dependent organizational deficiency such as failure to generalize the instructions of the task.

concern carries greater weight when test items are themselves abstract analogies because the result is a test containing instructions that can only be generalized to every item by those individuals who would be expected to perform well even in the absence of instructions.

I am *not* proposing that the ability to map dissimilar objects is a binary skill that a person either does or does not possess. As my brief mention of parsimony suggests, relevant productions are expected to characterize a range of skill levels within single cohorts. This much is conceded in the large grain-size of the productions specified above, which are too general to be translated into a production model that accurately simulates a single individual without making further assumptions. I have completely ignored the possibility of even higher-level organizational knowledge that may characterize very skilled problem solvers such as those who possess substantial *declarative* knowledge about the structure of analogies. Instead, the list of productions captures a category that includes distinct but functionally similar production lists of individuals who are similar only in the sense that they are capable of representing roles as unknowns. To complicate matters slightly, it is probable that more effective productions are accompanied by greater access to declarative roles and relations as a result of more frequent retrieval. I do not return to this possibility until the general discussion, but it is important to consider that a procedural change can be expected to have declarative consequences (and vice versa).

Mapping in Matrix Reasoning and the Cognition of Culture-Free Tests

The body of literature on matrix reasoning offers a glimpse at how procedural knowledge could enhance performance on tests with the largest Flynn effects. Matrix reasoning is a class of tests modeled after the well-known Ravens. Several characteristics of the Ravens make it ideal for studying the processes underlying variation in performance and how it relates to the Flynn effect. The test has one of the largest Flynn effects (Brouwers et al., 2009; Flynn, 1987), making it an ideal instrument for evaluating the present theory. Most importantly, items differ along simple dimensions such as number of rules.

The items on all matrix reasoning tests are organized in a similar manner such that rules must be identified from the interrelations of objects in an array to determine which response choice would best complete the array. Figures 1 through 5 are examples of matrix reasoning items. Each consists of a three-by-three array of abstract figures. One figure is absent on the lower right-hand

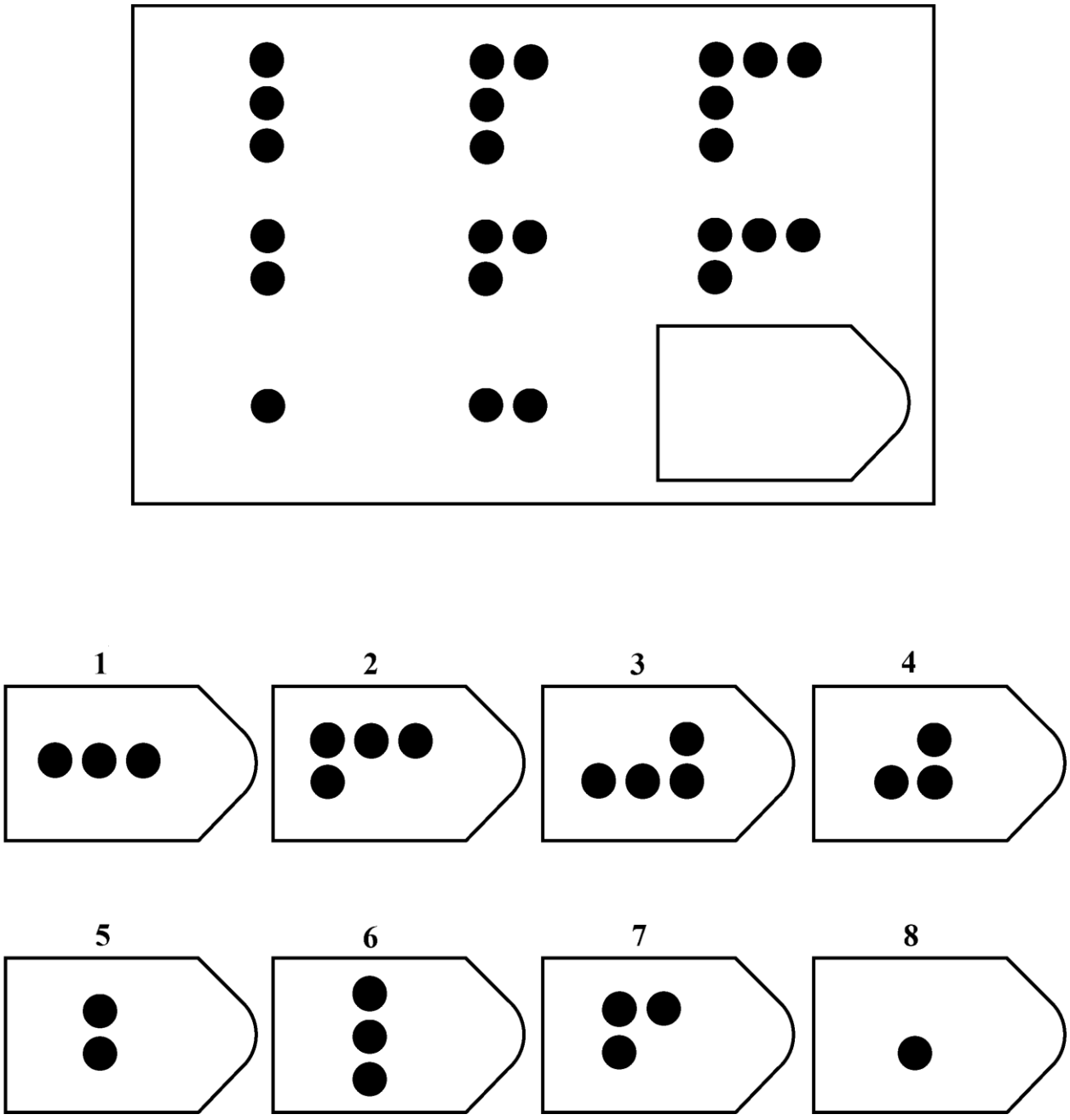

Figure 1. A simple item with quantitative-pairwise-comparison and constant-in-a row rules in Carpenter et al.'s (1990) taxonomy. Identical dots decrease in number from top to bottom in columns and increase in number from left to right in rows. Both rules are classified as level 1 because objects are synonymous with their roles (presence and placement within a figure).

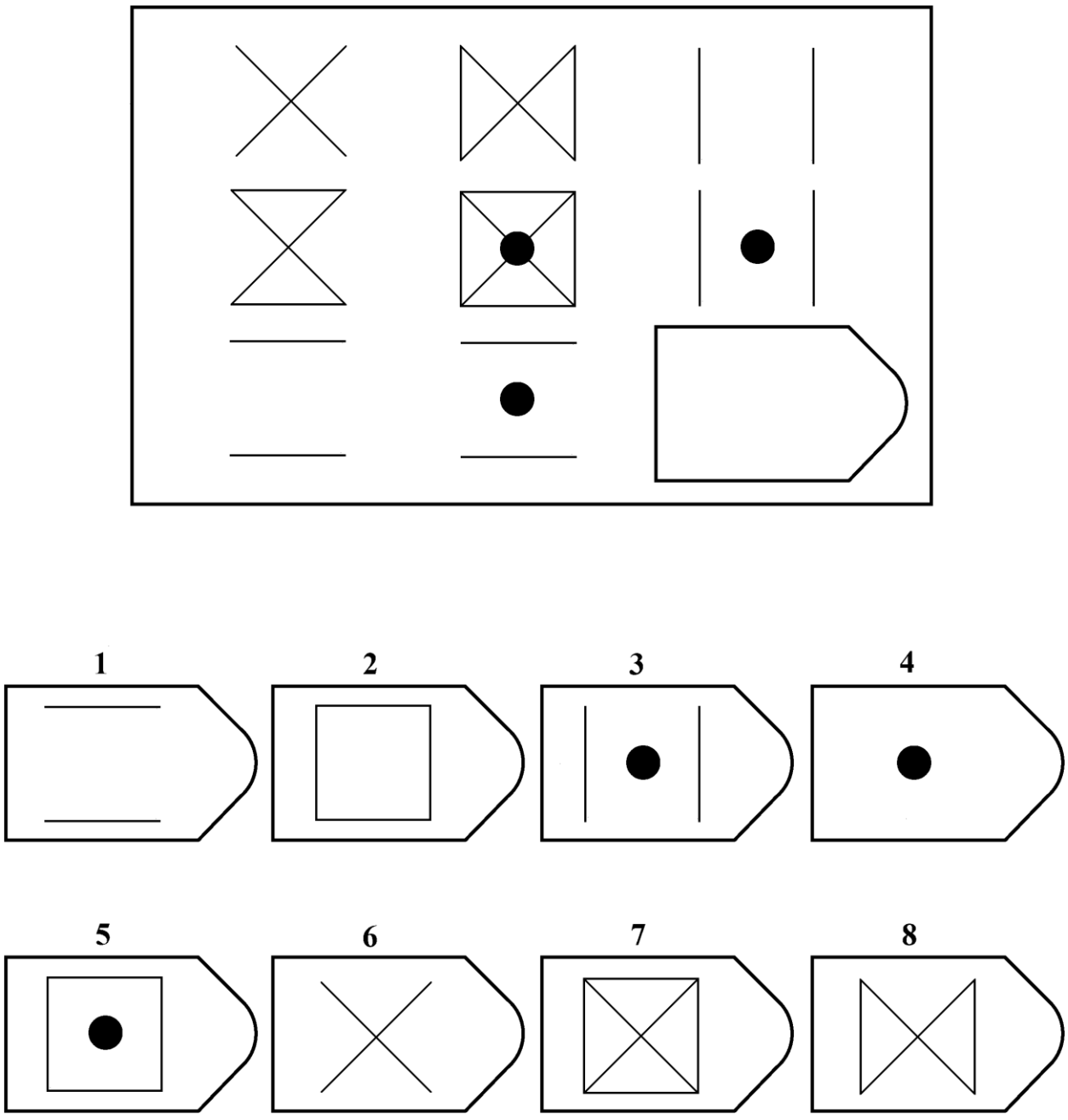

Figure 2. A figure-addition or subtraction and distribution-of-two-values item. As an addition/subtraction item, Figure 2 is classified as level 2 because objects with the same roles occupy the same location within rows or columns, but do not necessarily appear similar because some objects are absent.

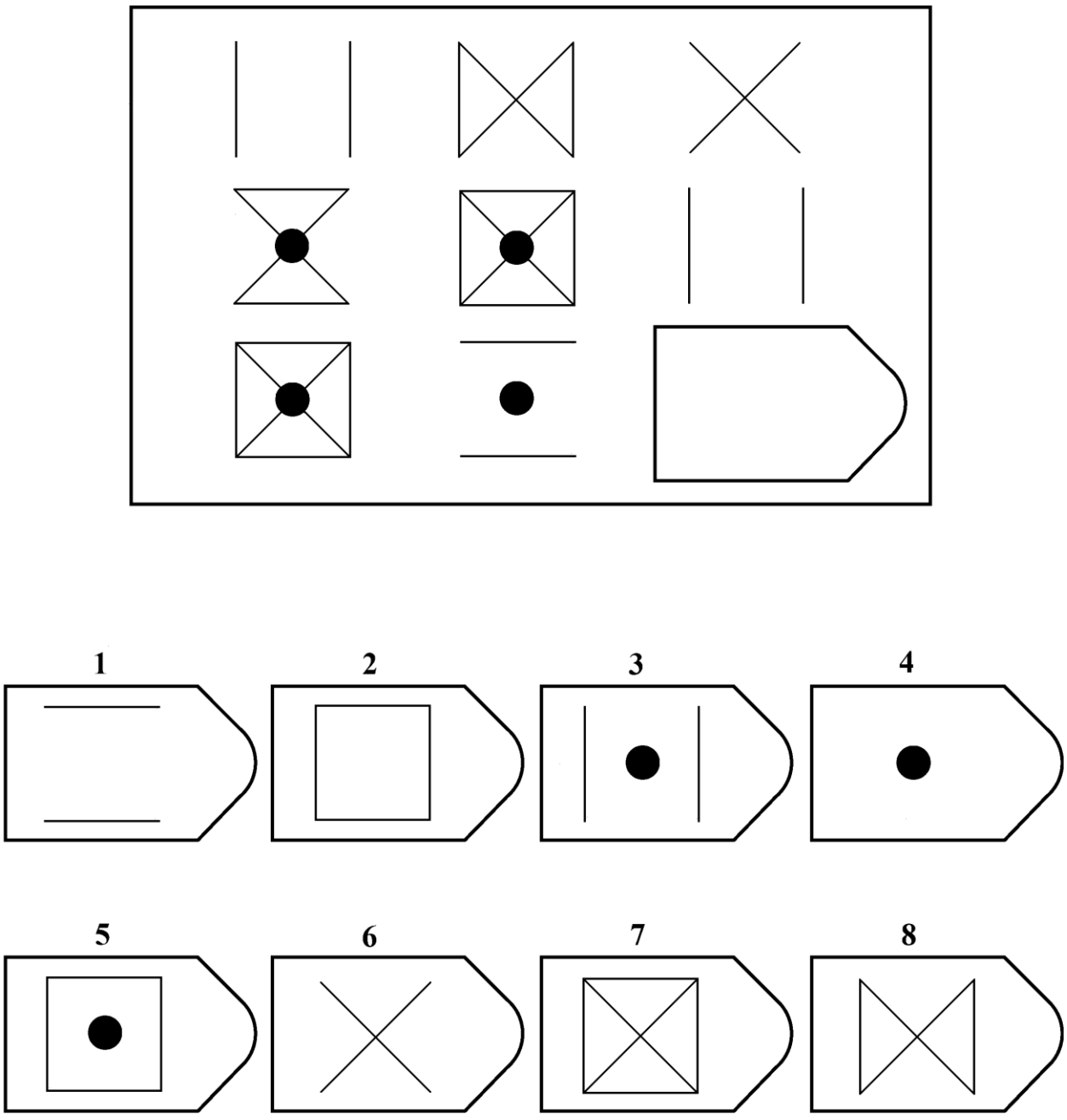

Figure 3. A modified version of Figure 2 that would be classified as the theoretical intermediate between level 2 and 3 (see Table 2). Unlike Figure 2, Figure 3 cannot be solved with ordinary addition or subtraction (using whole rows or columns) because objects with the same role (e.g., *addend, sum*) do not occupy the same location or appear similar across rows and columns. In other words, individual rows and columns must be added or subtracted separately.

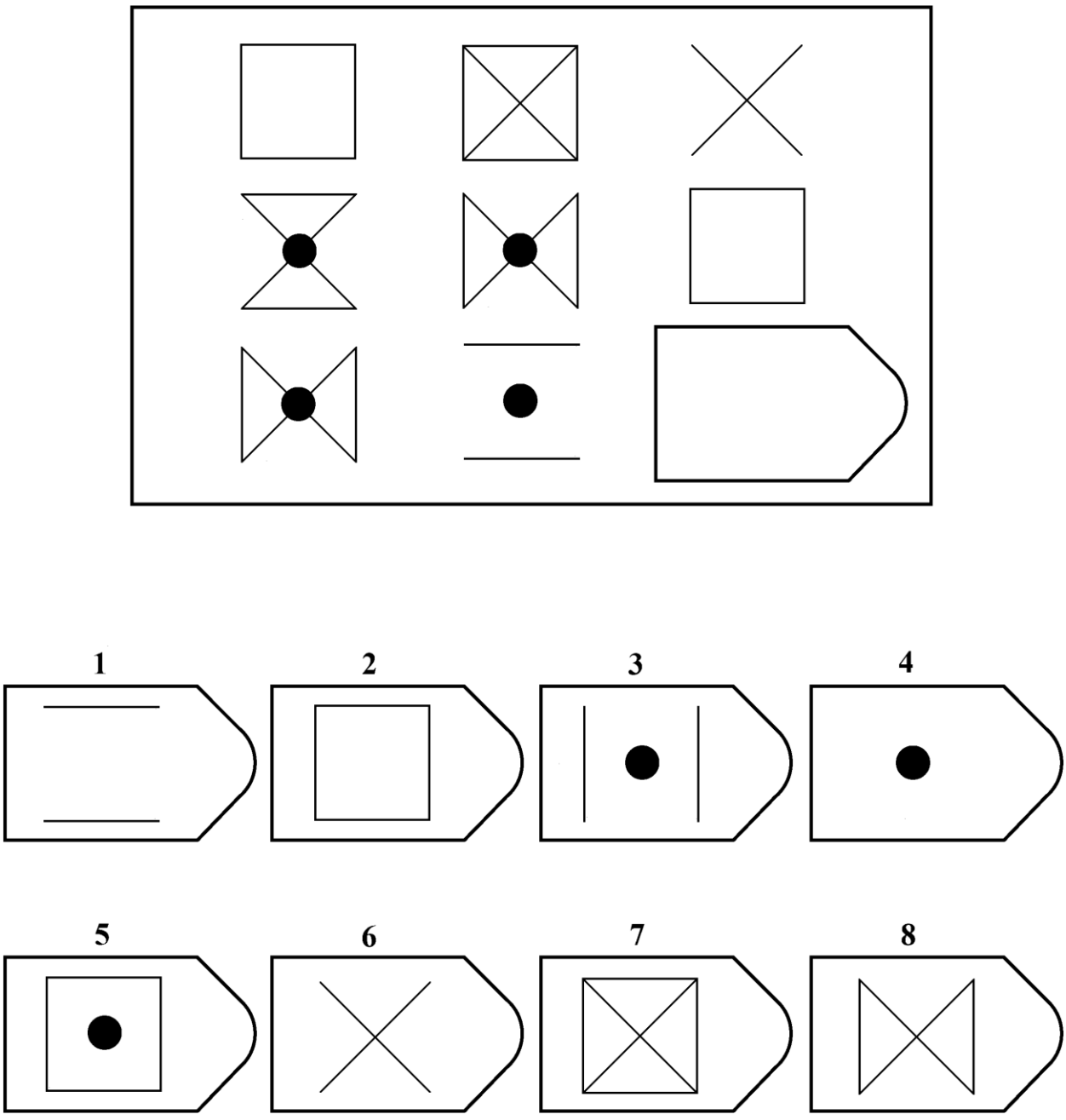

Figure 4. A modified distribution-of-two-values item that cannot be solved using an addition or subtraction rule. The item is classified as level 3 because the abstract role, *pair,* does not have the same appearance or placement in every row or column.

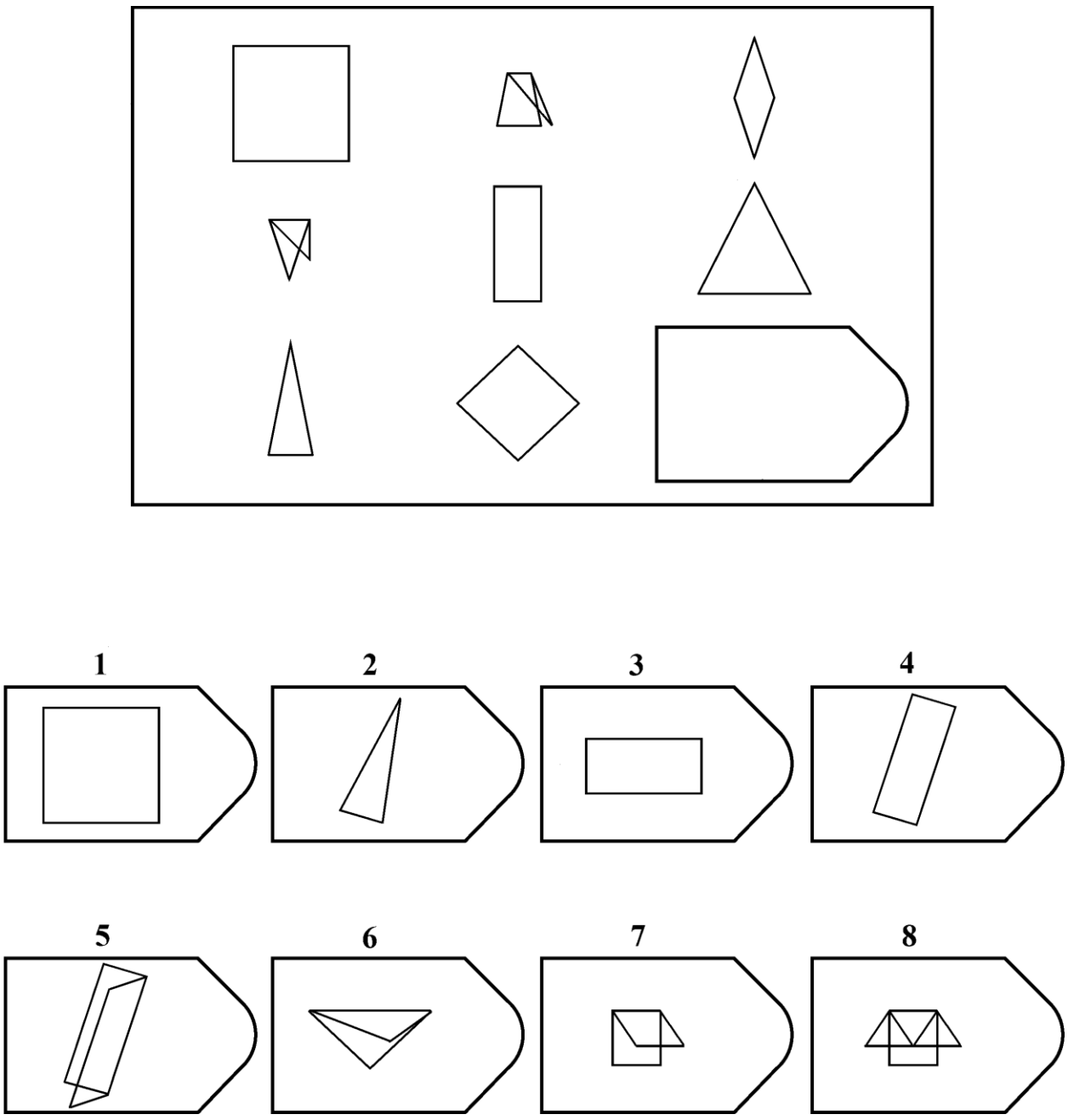

Figure 5. A distribution-of-three-values item. Corresponding objects are one of three shapes with wide and narrow versions, with or without a fold. The rules are classified as level 3, but most educated problem solvers are likely to represent the shapes of a given type (e.g., triangle) as similar.

side, and the goal is to determine which of the eight figures below best completes the array. Selecting the correct response requires inferring a rule that characterize each of the relationships among objects in the rows and the columns. Conceptually, either the rows or columns can be

considered analogs, but assuming for illustrative purposes that rows are analogs, a rule must be applied to the lowest row to determine which objects should be present in the missing figure. Corresponding objects for individual rules in matrix reasoning items vary in similarity from rule to rule. Carpenter, Just, and Shell's (1990) rule taxonomy is presented in Table 1.

Figure 1 is a simple item that can be solved by mapping similar objects. Dots, each of which is perceptually similar to the others, increase in quantity on the top from left to right. A second rule is that the quantity decreases on the side from top to bottom. Note that corresponding objects for a rule are present within single rows and columns. Every figure in the left column is one dot wide, every figure in the middle column is two dots wide, and every figure in the right column is three dots wide. Every figure in the top row is three dots tall, every figure in the middle row is two dots tall, and every figure in the bottom row is one dot tall. This neat spatial organization of quantities is not a necessary condition for a coherent item. A rule requiring one of each quantity (one, two, or three dots) in every row and column, regardless of location, would entail mapping the same physical objects but would require a more abstract rule. That is, the rule, *one dot on the left, two dots in the middle, and three dots on the right,* is less abstract than *one of each of the quantities, one, two, and three dots.*

Both rules of Figure 1 occupy the lowest level of dissimilarity (level 1) depicted in Table 2, which shows how rules must become more abstract and inclusive as dissimilarity of corresponding objects increases. The progression of dissimilarity is from level 1, where objects with corresponding roles have the same physical appearance, physical placement, and function; to level 2, where objects with corresponding roles have the same physical appearance *or* physical placement and function; to level 3, where objects with corresponding roles have only the same function, and not the same physical appearance or placement. The more abstract rule, *one of each of the quantities, one, two, and three dots,* occupies level 2, but would yield a correct solution when applied to a rule at level 1. Although at least one of the levels in the table is applicable to any rule of any item in the Ravens, the levels are applied to Figures 2, 3, and 4 in particular to illustrate the ordinal nature of increasing abstractness as a function of minor changes in item properties. Figure 2 is considered an addition-subtraction item in Carpenter et al.'s (1990)

Table 1

Carpenter et al.'s (1990) Taxonomy

Rule	Taxonomy
Constant-in-a-row	The same value occurs throughout a row, but changes down a column (Figure 1).
Quantitative-pairwise-progression	A quantitative increment or decrement occurs between adjacent entries in an attribute such as size, position, or number (Figure 1).
Figure-addition or subtraction	A figure from one column is added to (juxtaposed or superimposed) or subtracted from another figure to produce a third (Figure 2).
Distribution-of-three-values	Three values from a categorical attribute (such as figure type) are distributed through a row (Figure 5).
Distribution-of-two-values	Two values from a categorical attribute are distributed through a row; the third variable is null (Figures 2, 3, and 4).

Note. Table is reproduced verbatim (p. 408) except for examples. Abstractness and dissimilarity of objects tends to increase in ascending order.

taxonomy (see Table 1) because objects in the middle column and middle row are the concatenation of objects in the other two columns or rows. Thus, Figure 2 can be solved with the relatively concrete rule, *right and left appear in the middle* (level 2). Figure 3 is a simple modification of Figure 2 that requires a slightly more abstract application of addition or subtraction as it applies within single rows or columns, *one plus another equals the third* (the intermediate level in Table 2). By rearranging the objects in Figure 2 and 3, it is possible to create an item with the most abstract rule in Carpenter et al.'s (1990) taxonomy. Figure 4 is a distribution-of-two-values item, or as it is presented in Table 2, *two of a kind.* The role, *pair,* does not have similar physical appearance or placement in every row and column.

The last example, Figure 5, illustrates the importance of the distinction between representation and physical properties of an item. Even though the three-sided shapes in each

row differ somewhat in appearance from one another, they are clearly apprehended by the reader as instances of the same role, *triangle*. Because this role may not be as universal and obvious as it seems, I commit to a formalist perspective for present purposes, whereby Figure 5 entails mapping objects that are dissimilar. They are dissimilar because corresponding objects differ in appearance and do not occupy the same rows and columns as do the corresponding objects in Figure 1. A relation such as *basic shape* (level 3) must be generated if the common roles of *triangle, square,* and *diamond* are not retrieved automatically (in which case the rule would be classified more accurately as level 2). The purpose of this example is to show that there is no operational definition of dissimilarity that can be expected to apply to every population for the same reason that it is necessary to invoke procedural knowledge as a mechanism of transfer. It is the representation, not the superficial content, that makes one object similar to another.

As Table 2 suggests, evaluating the abstractness of rules in terms of object roles results in abstractness rankings that are comparable to Carpenter et al.'s (1990) rankings of rule difficulty (ascending order of Table 1). Abstractness necessarily covaries with dissimilarity of objects because more abstract rules refer to features of objects that differ from initial representations. To reiterate, abstract rules subsume concrete rules such that concrete rules often do not generalize beyond single objects whereas abstract rules may generalize to entirely different analogies on different tests. Every Ravens item can be solved by mapping objects at the third level of abstractness or lower.

An analysis of matrix reasoning requires considering two potential sources of variation: variation in item properties and variation in test-takers. As Vigneau, Caissie, and Bors (2006) point out, these sources need not comply with one another as the rank order of item difficulty need not be the same for every individual in a test sample. I consider variation in item properties within cohort to establish that items with rules containing dissimilar objects are more difficult than items with rules containing only similar objects. Although the greater difficulty of rules with dissimilar objects may seem obvious, it cannot be taken for granted. It is conceivable that other item properties such as the number of rules that must be inferred are sufficient to account for all the variation in item difficulty. Establishing that dissimilarity of objects is a source of variation in item difficulty within one cohort is the first step toward establishing that the ability to map dissimilar objects varies between cohorts.

Table 2

Relationship between Dissimilarity of Objects and Abstractness of Rules for Figures 2, 3, and 4

Level of dissimilarity	Similarities of objects with same role	Example of relation	Example of role	Application to Figures 2, 3, and 4
1	Physical appearance, physical placement, and function	*Vertical lines on right and middle*	*Vertical lines*	Incorrect response to Figures 2, 3, and 4
2	Physical appearance *or* physical placement and function	*Right plus left equals middle*	*Right figure* (any objects)	Correct response to Figure 2; incorrect response to Figures 3 and 4
Theoretical intermediate[a]	Only function (but dependent upon physical organization of objects within an analog)	*One plus another equals the third*	*Addend 1* (any figure (any objects))	Correct response to Figures 2 and 3; incorrect response to Figure 4
3	Only function (but indifferent to the physical organization of objects within an analog)	*Two of a kind*	*Pair* (any class of object (any figure (any object)))	Correct response to Figures 2, 3, and 4

Note. Parentheses contain the more concrete (less generalizable) categories that are subsumed by a role. The representation of objects at every level subsumes the representation at lower levels. Similar physical placement means placement within the same row or column.

[a] This level is not represented by the Ravens items used in Studies 1 and 2.

Working memory has received considerable attention in regards to both forms of variation, as a resource demanded by items in the case of item difficulty, and as a resource invested in overcoming this demand in the case of individual differences. It is noteworthy that authors of virtually every study I review below suggest working memory as the explanation of their findings. There are two basic accounts of how working memory relates to matrix reasoning, one of which is straightforward enough to be considered alongside item properties. This is the claim that items with more rules demand more working memory (e.g., Carpenter et al., 1990). A second more complex claim is that rules containing dissimilar objects place greater demands on working memory. Although this claim is not necessarily at odds with the present theory—it is possible for dissimilarity of objects to moderate difficulty for more than one reason—I argue below that it is without support unless one defines working memory as something that is logically indistinguishable from procedural knowledge. This conflict cannot be resolved without considering the causes of individual differences, in particular, the work of Carpenter et al. (1990). I begin by showing that items with more rules and items that require mapping dissimilar objects are more difficult to solve. Following this brief review, Study 1 assess whether the pass rates of items with dissimilar objects have increased disproportionately as the theory predicts. Only after Study 1 do I consider evidence that differences in the ability to map dissimilar objects are caused primarily by differences in knowledge.

The studies reviewed in this section report data collected from participants born recently enough (most after 1970) to be considered samples of the same large, recent cohort. Within these studies, working memory load has received the most attention as a determinant of item difficulty. To summarize, Carpenter et al. (1990) concluded that number of rules, number of rule tokens, and type of rule are associated with item difficulty as more rules and rule tokens are associated with lower pass rates. Hornke and Habon (1986) and Embretson (1998) were able to generate items that correlated highly with the Ravens, and were able to accurately predict the difficulty of specific items by considering number of rules and rule tokens.

However, there is also evidence that dissimilarity of objects is a source of variation. Following up on Carpenter et al.'s (1990) claim that rules with dissimilar objects are more difficult to induce because they demand more working memory, Embretson (1998) and Primi (2002) constructed experimental items and found that items containing these rules are less likely to be solved. Embretson's (1998) analysis revealed lower probabilities of solution when

corresponding objects are not located in same rows and columns, while Primi's (2002) study revealed lower probabilities of solution when corresponding objects are perceptually dissimilar. In a similar study, Meo, Roberts, and Marucci (2007) manipulated object familiarity by constructing items with common letters and novel letter-like symbols that were isometric (in terms of relations between objects) to those from the Raven's Standard Progressive Matrices and Advanced Progressive Matrices. According to the present theory, these new items should have been easier than the original Ravens items because the investigators essentially accomplished the mapping for participants by labeling the corresponding objects with either familiar letters or easily dissociable symbols. Solving Figure 5 is difficult, in part, because the corresponding elements are perceptually dissimilar. Such an item, based on Meo et al.'s (2007) encoding scheme, would use a single letter or symbol to represent the corresponding objects of each shape (e.g., all triangles are "A"), thus allowing mapping to occur by simply "reading" the figures. The theory also predicts greater difficulty of letter-like symbols compared to ordinary letters for a similar reason. Objects that can be recognized as similar are more likely to be mapped to one another. The superficial properties of familiar letters are a single, meaningful chunk, whereas the superficial properties of letter-like symbols may represent at least several depending on how familiar they are to ordinary letters. According to the theory, familiar letters should be more easily recognized as similar and should be mapped with greater facility. Meo et al.'s (2007) findings were consistent with these assumptions as the Ravens items were the most difficult, followed by letter-like symbols, and then ordinary letters. These studies are the primary empirical justification for the levels of dissimilarity (or conversely, levels of abstraction) presented in Table 2.

To summarize what I have presented thus far, the literature on matrix reasoning suggests two primary within-cohort causes of variation in item difficulty. It is well-recognized that items with a greater number of rules are more difficult, but additional research suggests that items with rules containing dissimilar objects are also more difficult. According to the theory, these latter items in particular should have become easier to solve over time, at least in the short term. That is, if the Flynn effect has been caused by improvements in the ability to map dissimilar objects, recent gains should be greatest on items with dissimilar objects.

CHAPTER THREE

STUDY 1: PREDICTING GAINS IN ITEM-SPECIFIC PASS RATES

The primary goal of Study 1 is to compare item-specific pass rates in two test samples from two time periods with virtually identical overall performance. The primary prediction is that pass rates in the more recent sample will be concentrated among items with rules containing dissimilar objects. The earlier pass rates were collected roughly four decades earlier than the later pass rates. Although pass rates do not have the resolution of individual item responses, the comparison is an excellent test because the mean pass rates (averaged across items) of the samples are virtually identical, presumably, because the earlier sample (Forbes, 1964) was selected with a goal of establishing discriminabilities for difficult items, which would have required a high proportion of participants who scored high relative to their cohort.

Forbes' (1964) "highly selected" (p. 223) sample was chosen to represent the highest range of ability targeted by the Ravens for the revised 1962 version of the test. For example, the sample included some unspecified proportion of participants with aspirations of becoming engineers (Forbes, 1964). The young adults and late adolescents were born around or shortly after 1940 and tested around 1961 (cohort 1). A modern sample (cohort 2) is comprised of three contemporary datasets (Mitchum & Kelley, 2010; Unsworth & Engle, 2005; Vigneau & Bors, 2005). The two cohorts are roughly matched on overall performance, which helps to rule out alternative explanations. Factors that contribute to higher scores within cohorts, but have not changed over time, should favor cohort 1 because participants in this sample are likely to be more proficient at tests like the Ravens relative to their own population. Consequently, item-specific differences observed in favor of cohort 2 are unlikely to be attributable to any unknown variable that is irrelevant to the Flynn effect. It is predicted that number of rules (meaning number of tokens and not number of types[6]) will correlate highly with pass rates within cohorts,

[6] This variable has little immediate relevance to testing predictions. For the benefit of interested readers, the correlation between the number of types of rules and outcome variables is low to non-existent in each set of pass rates.

but because the theory does not predict that working memory has improved, it is predicted that number of rules will not correlate with magnitude of changes in pass rate between cohorts.

Method

Datasets. Item-specific reliabilities for the Ravens are low (Bors & Vigneau, 2001), meaning that large datasets are needed to achieve sufficient mean estimates to test predictions about changes in pass rate. Although cohort 1 is already sufficiently large (n = 2,256), maximizing the reliability of pass rates for cohort 2 required combining pass rates from multiple datasets. A literature search for studies reporting item-specific pass rates yielded Vigneau and Bors (2005, n = 506), and a study by Unsworth and Engle (2005, n = 160). Pass rates of the samples used by Mitchum and Kelley (2010) were also added (n = 117). Administration conditions were not identical across each study (see original articles for details), and birth years of the 783 participants span more than a decade. Nonetheless, overall performance was comparable and item-specific pass rates were internally consistent (α = .92)[7] as correlations of pass rates between any two samples exceed r = .9. Roughly equal pass rates and uniform item difficulty over a period of about a decade for cohort 2 is consistent with evidence that the Flynn effect has ceased in some countries that showed the earliest gains (e.g., Cotton, Kiely, Crewther, Thomson, Laycock, & Crewther, 2005; Lynn & Harvey, 2008; Sundet, Barlaug & Torjussen, 2004). Mean pass rate for cohort 1 (M = .60, SD = .29) and cohort 2 (M = .59, SD = .26) are closely matched (d = .04). Figure 6 shows pass rates as a function of item number for both cohorts.

Although descriptives are not provided by every study, the information available suggests participants in Cohorts 1 and 2 were comparable in age at the time of testing with mean ages of about 20 years. Cohort 1 is comprised of young adults and some late-adolescents. Cohort 2 consists entirely of undergraduates from psychology department participant pools. Sex is confounded with sample as cohort 1 is primarily male (at least 66 percent based on information given), and cohort 2 is primarily female (roughly 60 percent). There is tentative evidence for a minor male advantage on the Ravens (Mackintosh & Bennett, 2005; Vigneau & Bors, 2008), but

[7] Applied to pass rates of groups instead of accuracies of individuals, this value treats within-sample, between-subjects variation as error. This is not a serious concern because pass rate, not individual accuracy, is the unit of analysis. For the record, comparing each of the datasets in isolation to cohort 1 (the largest sample) reveals the same pattern of findings as those reported below with comparable effect sizes in each case.

this advantage is not robust (Vigneau & Bors, 2008), and is probably too small to pose a concern given that the predicted advantage for cohort 2 on more abstract items should exceed the effect size of sex within cohorts if the theory is accurate. Cohort 1 data were collected in the United Kingdom, and cohort 2 data were collected in the United States and Canada. Both regions witnessed large Flynn effects (Flynn, 1987), and there is no obvious reason other than the Flynn effect itself to expect contemporary North American young adults to be better at solving abstract items than their transatlantic counterparts were fifty years ago.

Figure 6. Comparison of item-specific pass rates of cohorts 1 and 2 for Study 1. Error bars represent +/–1 standard error.

Item classifications. Carpenter et al.'s (1990; p. 431) classifications were used to assign number of rules to Ravens items. Carpenter et al. (1990) do not report numbers of rules for 11 of the 36 items either because the authors were unable to use the item in their analysis (n = 9) or because the item cannot be classified according to their taxonomy (n = 2). To maximize the number of observations available for analysis, I assigned numbers of rules to the 9 compatible items using Carpenter et al.'s (1990) taxonomy, and assigned numbers of rules to the remaining

two items using novel rules. The decision to include items that were not classified by Carpenter et al. (1990) does not alter the pattern of results reported below.

Developing perfect criteria for classifying the dissimilarity of corresponding objects requires nothing less than knowledge of how participants represent objects. Because obtaining this information from pass rates is impossible, simple criteria were used to optimize the trade-off between simplicity and cognitive plausibility. The simplest and least controversial solution, compatible with the studies reviewed above, is to classify similarity with respect to relative placement of corresponding objects (whether or not they are found in the same rows or columns), and appearance of corresponding objects (whether or not they appear identical to one another). However, several minor exceptions are needed to improve the cognitive plausibility of appearance as a criterion for similarity. Specifically, corresponding objects that differ in size, but remain otherwise perceptually identical (this includes lengths of single lines, e.g., item 10), and identical shading patterns (which may appear on different shapes; e.g., item 21) are considered similar. As noted, perceptually non-identical shapes such as the three triangles in Figure 5 are considered dissimilar based on the present criteria. I return to this issue in the general discussion.

Each rule of every item was classified as one of the three levels of dissimilarity presented in Table 2. Rules were assigned to the lowest (most similar) level that is sufficient for correct mapping of objects. Because the theory makes no assumptions about whether participants represent columns or rows as analogs, the lowest level of dissimilarity compatible with solution was established by considering similarities from row to row and column to column. In accordance with Table 2, rules were classified as level 1 if corresponding objects are similar in appearance and occupy the same figure within their respective rows or columns. Rules were classified as level 2 if corresponding objects are similar in appearance *or* occupy the same figure within their respective rows or columns. Finally, rules were classified as level 3 if corresponding objects are dissimilar in appearance and occupy a different figure within their respective rows or columns. As expected, the classifications overlap considerably (about $r = .6$) with a variable representing Carpenter et al.'s (1990) ranking of rules by difficulty (see Table 1).[8]

Results and Discussion

[8] I do not include Carpenter et al.'s (1990) difficulty of rules as an additional variable because it is virtually identical to the variance shared between dissimilarity and number of rules. For this reason, it is correlated with pass rate and change in pass rate, but never shares unique variance with either of these outcome variables.

Given the limited number of observations (1 for each of 36 items per sample), sophisticated regression models are unjustified for examining relationships between dissimilarity, number of rules, and the outcome variables of pass rate and change in pass rate. Because number of rules and dissimilarity are correlated (r = .42), linear regression is used, but primarily to obtain partial correlations representing unique variance shared between either predictor variable and pass rates or change in pass rates.

Consistent with previous research (Carpenter et al., 1990; Embretson, 1998; Primi, 2002), it was found that number of rules and dissimilarity both correlate with item-specific pass rates in both cohorts. In regression models, the two variables account for nearly two-thirds of the variance in pass rate in cohort 1 (R^2 = .66) and cohort 2 (R^2 = .61). Effect sizes for number of rules are large in cohort 1 (r = –.68, 95% CI [–.82, –.45]) and cohort 2 (r = –.68, 95% CI [–.82, –.45]), implying that the working memory load of additional rules raises item difficulty. The effects sizes for dissimilarity are comparable to those of number of rules in cohort 1 (r = –.69, 95% CI [–.83, –.47]) and cohort 2 (r = –.64, 95% CI [–.80, –.40]). These results are consistent with findings of Carpenter et al. (1990), Embretson (1998), and Primi (2002). Because number of rules correlates with dissimilarity, it is informative to consider the unique variance that either predictor shares with pass rates within either cohort. Partial correlations reveal that both number of rules (cohort 1: r = –.59, 95% CI [–.77, –.32]; cohort 2: r = –.59, 95% CI [–.77, –.32]) and dissimilarity (cohort 1: r = –.61, 95% CI [–.78, –.35], cohort 2: r = –.53, 95% CI [–.73, –.24]) are strong independent predictors of pass rate in both cohorts. These within-cohort results concur with previous research and verify the assumption that level of dissimilarity contributes to item difficulty in multiple cohorts.

To test the prediction that pass rates increased more on items with dissimilar objects, gains in item-specific pass rates from cohort 1 to cohort 2 were calculated by subtracting pass item-specific pass rates of cohort 1 from those of cohort 2. These changes in item-specific pass rates approximate a normal distribution (kurtosis and skewness are well within the range of +/–1; K-S Z = .52, p = .95) and are treated as a continuous dependent variable in the following analysis. Regression revealed positive relationships between changes in item-specific pass rate and both number of rules and dissimilarity. However, the effect size for number of rules (r = .29, 95% CI [–.04, .56]) disappears when unique variance is isolated (r = –.07, 95% CI [–.39, .26]). In contrast, the effect size of dissimilarity is larger (r = .58, 95% CI [.31, .76]) and remains virtually

unchanged when unique variance is isolated (r = .53, 95% CI [.24, .73]). As predicted, dissimilarity is positively associated with changes in pass rates, as cohort 2 gains were concentrated in items with dissimilar corresponding objects. Although number of rules may also be associated with change in pass rate, whatever association there is appears to be a consequence of items with more rules also tending to have rules with dissimilar objects.

In accordance with previous work, Study 1 shows that number of rules and dissimilarity of objects are sources of variation in item-specific pass rates from at least two large samples collected nearly half a century apart. As predicted, cohort-related gains in pass rates are associated with dissimilarity of objects, but not number of rules. These gains in performance on the test with the largest Flynn effect are consistent with the assumption that, over about a forty-year period, young adults became better at mapping dissimilar objects. While results of Study 1 are constrained by the resolution of group-level observations, the comparison is based on over 3,000 participants and reveals the expected difference in pass rates for items with rules containing dissimilar objects despite equivalence in overall scores of the two cohorts. Results of Study 1 could stand alone as compelling evidence that the Flynn effect reflects changes in the mapping of dissimilar objects.

Nonetheless, the effect size of around r = .53 has a very large confidence interval and should be interpreted cautiously. As noted previously, predicting how participants will represent objects is very difficult, and reasonable arguments could be made for altering some of the item classifications in light of cultural factors as well as theories of perception and categorization. Moreover, it should not be forgotten that differences between pass rates of items are not identical to differences between persons, much less the cognition of individual persons. If the theory is accurate, a shift in focus to variation in persons should reveal that individual differences in scores are caused by individual differences in mapping of dissimilar objects.

CHAPTER FOUR

SOURCES OF VARIATION IN SCORES: AN ALTERNATIVE INTERPRETATION OF CARPENTER, JUST, AND SHELL (1990)

Few studies have considered whether individual differences in scores are caused by individual differences in the ability to map dissimilar objects. One exception is a study by Schiano, Cooper, Glaser, and Zhang (1989), which showed that high performers sort geometric analogies problems into categories based on deep similarities, while low performers were more likely to categorize based on mere appearance of objects. This finding is consistent with the assumption that high performers represent objects according to functional relations, whereas low performers are more likely to simply map similar objects. In contrast to Schiano et al. (1989), most investigators have tended to focus on differences in working memory. My goal in this section, without going so far as to dismiss working memory as a causal variable, is to show that much of this focus has been misdirected. In particular, the assumption that differences in working memory cause differences in the ability to map dissimilar objects (Carpenter et al., 1990; Primi, 2002) is based on a somewhat speculative interpretation of a straightforward finding.

Carpenter et al.'s (1990) models of Ravens performance are a monumental contribution to explaining the cognition of individual differences. Although the argument that follows is somewhat critical, it is aimed at a very specific claim about working memory, and does not challenge the substance of the models. While acknowledging that these models are pillars of the present work, I suggest that Carpenter et al.'s (1990) interpretation of them muddles the concept of working memory, and that the models warrant a less speculative interpretation as evidence that differences in the ability to map dissimilar objects are caused by differences in procedural knowledge.

The two production models were implemented in CAPS (Concurrent, Activation-Based Production System). High performers (the BETTERAVEN model) and lower performers (the FAIRAVEN model) were simulated, based on eye movement and verbal protocol data. Carpenter et al. (1990) simulated differences in the ability to manage multiple goals by giving

BETTERAVEN "a goal monitor that sets strategic and tactical goals, monitors progress toward them, and adjusts the goals if necessary. In addition, the control structure of BETTERAVEN, as governed by the goal monitor, is somewhat changed" (p. 419). The goal monitor for BETTERAVEN was instantiated by adding 15 productions to FAIRAVEN. The additional goal-monitoring productions made BETTERAVEN better at identifying rules and helped to simulate differences in accuracy between the models similar to those found between individuals high and low in working memory.

Whether these differences between the models merit a working memory interpretation depends on whether one considers working memory a literal quantity or an emergent property of multiple causes that happens to lend itself to interpretation as a quantity in a metaphorical sense. Constructs believed to represent literal quantities are most meaningfully simulated by varying a parameter such as source activation (the *W* parameter) in ACT-R (see Anderson, Reder, & Lebiere, 1996; Daily, Lovett, & Reder, 2001). Carpenter et al. (1990) varied productions, but seem to suggest that differences in productions are the *result* of differences in working memory because working memory facilitates goal-management. "One of the main distinctions between higher scoring subjects and lower scoring subjects was the ability of the better subjects to successfully generate and manage their problem-solving goals in working memory" (p. 428). The problem is that formalizing a single construct simultaneously as both a cause and an emergent property of this same cause introduces a logical dependency (e.g., see Boag, 2011; Michell, 2011) that defeats the purpose of the construct as an explanation.

The models have a simpler interpretation as evidence that mapping of dissimilar objects requires more advanced procedural knowledge. FAIRAVEN follows a systematic procedure whereby encoding of perceptual features (represented in hand-coded information that is interpretable to the model) is followed by determination of correspondence between figures, and then determination of correspondence of objects from different figures. The next step is the critical rule-identification phase that the authors term "conceptual analysis" (p. 417). Importantly, FAIRAVEN is given knowledge of four of the five rule types identified by Carpenter et al. (1990). The model cannot handle distribution-of-two-values rules that require mapping of dissimilar objects. Consequently, FAIRAVEN assumes that objects corresponded to one another only if verbal protocols revealed that they were typically given the same name by participants (e.g., *line* or *circle*). This characteristic of FAIRAVEN captures the difficulty low

performers have with inducing the distribution-of-two-values rule and Primi's (2002) finding that they are less able to map perceptually dissimilar objects. In contrast, BETTERAVEN's additional goal monitor allows it to test other rules when the mapping of matching names does not successfully elicit a rule. This ability of BETTERAVEN approximates the process of testing new rules instead of relying on a default procedure that maps on the basis of similarity. This process leads to consideration of new rules even when figures have dissimilar objects. Although neither model generates rules, it is clear that BETTERAVEN is better at mapping dissimilar objects because of its better and more flexible productions, irrespective of whether these productions are caused by greater working memory.

While I insist that empirical findings cannot inform the *meaning* of constructs (as previously stated, concepts, including psychological constructs, are psychological, *not* empirical, entities), it is noteworthy that a widely accepted operational definition of working memory reveals findings compatible with the distinction I have drawn between procedural knowledge and working memory. Unsworth and Engle (2005) examined correlations between the well-known Operation span working memory test, and each of the 36 Ravens items. Although Operation span correlated with overall performance (r = .34), the authors found no increase in the relationship between Operation span and accuracy on abstract items that require mapping of dissimilar objects as a working memory interpretation of Carpenter et al.'s (1990) work would predict. In fact, the fourth quartile of the test, featuring the most abstract items, was the only quartile in which score was uncorrelated with Operation span. Further analysis revealed an expected main effect of rule type as distribution-of-two-values and distribution-of-three-values rules in particular had higher error rates. However, Operation span did not interact with rule type, as would be expected if inducing abstract rules requires more working memory. Although a null finding can result by chance, a graph of correlations between span and Ravens accuracy as a function of item (Unsworth & Engle, 2005, p. 73) reveals no signs of a trend, suggesting that any real but undetected relationship is likely to be small. Just as interestingly, the relationship between Operation span and accuracy did not increase with the number of rules or tokens in an item as would be expected if greater working memory capacity enhances performance by making it possible to preserve old information during execution of new processes.

Finally, and as I have already suggested, the relative effect of one variable on performance of some criterion task is dependent on the extent to which other variables vary, which is why there

is no reason to assume that a variable that limits performance in one population also limits performance in others. For example, differences in procedural knowledge would not be expected to cause substantial variation in scores in a population where everyone practices tests, but procedural knowledge would remain a major determinant of absolute performance. Carpenter et al. (1990) seemed to acknowledge this, both distinguishing what is common to every person from what causes them to differ (p. 429), and cautioning that "low-scoring subjects sometimes use very different processes on the Raven test, which could obscure the relationship between Raven test performance and working memory for such individuals" (p. 427). The major difference between Carpenter et al.'s (1990) interpretation and my own appears to be a distinction between conceiving of between-subjects variation as an *actuality* versus a *potentiality*. At least in the context of the Flynn effect, I suggest that the latter conception merits consideration.

To summarize, Carpenter et al.'s (1990) models are, by all accounts, evidence that differences in procedural knowledge cause differences in the ability to map dissimilar objects, which in turn cause differences in matrix reasoning scores. The further claim that this procedural knowledge is determined by working memory is not supported by any direct evidence unless the meaning of working memory is incorporated into procedural knowledge. The principal question to be answered is whether this procedural knowledge varies between cohorts.

CHAPTER FIVE

STUDY 2: OBSERVING BETWEEN-COHORT VARIATION IN PROCEDURAL KNOWLEDGE

Study 2 uses a think-aloud methodology to test the prediction that individuals from a more recent cohort map dissimilar objects more effectively than individuals from an earlier cohort. A cross-sectional sample of participants from two cohorts (i.e., younger and older adults) separated by roughly 50 years complete the Ravens while thinking aloud (Ericsson & Simon, 1993) or remaining silent. Participants solve the items in either their standard ascending order or a random order to minimize any bias introduced by practice effects. Previous work reveals substantial improvement on the Ravens after even one testing session (Bors & Vigneau, 2001), which implies that problem solving on the later, more critical items is likely to be affected by practice on earlier items. For present purposes, preserving the psychometric properties of the test is less important than observing the ability of participants from both cohorts to map dissimilar objects in the absence of direct practice on early items.

The study relies on a cross-sectional sample of older and younger adults, which is less ideal than cross-cohort groups matched for age. However, this confound of age with cohort is mitigated by two considerations. First, most studies examining effects of biological aging on cognition use cross-sectional data and there is no reason to assume these studies are any less susceptible to the same confound. Although this does not excuse the present confound, it does provide a precedent for attributing a predicted effect to the cause that elicited predictions. The present predictions are distinct from those made by biological theories of aging. The relatively low performance of older adults on tests like the Ravens is generally attributed to one or more individual-difference constructs such as working memory capacity (Babcock, 1994; Viskontas, Holyoak, & Knowlton, 2005; Viskontas, Morrison, Holyoak, Hummel, & Knowlton, 2004). The predicted finding that younger adults map disproportionately better on items with dissimilar objects cannot be attributed to any age-related mechanisms without challenging the primary finding of Study 1, which revealed that a group of young adults who are now of comparable age to older adults in the present study performed relatively poorly on items requiring mapping of dissimilar objects.

Testing Process Theories with Think-Aloud Verbal Reports

Testing theories of the processes underlying complex tasks requires information beyond what can be inferred from accuracy and solution time data (Ericsson & Simon, 1980, Fox, Ericsson, & Best, 2011). Think-aloud verbal reports can be used to test a theory of the sequence of processes that must be executed to perform a task accurately, as revealed by a task analysis (Ericsson & Simon, 1980, 1993). Participants are asked to verbalize their "inner voice" while solving a problem, and the verbalizations recorded during the task are encoded according to criteria determined by a task analysis. According to Ericsson and Simon's (1980, 1993) theory, think-aloud verbalizations reflect the contents of working memory at the time of expression.

Ericsson and Simon (1980; 1993) argue that equal performance of think-aloud and silent conditions on a difficult task implies validity because there are relatively few ways to perform a task correctly (compared to incorrectly), and the working memory source of verbalizations is limited. Consequently, a participant who successfully executes a task that is challenging (to that participant) is unable to generate an invalid verbal trace of task execution. Unlike explanatory verbal reporting where participants are asked to explain what they are doing, or directed verbal reporting where participants are asked to verbalize specific information, think-aloud is generally non-reactive under the conditions specified by Ericsson and Simon (Ericsson & Simon, 1993; Fox et al., 2011). A meta-analysis by Fox et al. (2011) of the effects of verbalization that included 47 think-aloud comparisons, revealed an effect of think-aloud no greater than would be expected by chance ($r = -.03$). However, the issue of reactivity is especially relevant to the present study because Fox and Charness (2010) found evidence that thinking aloud improved older adult scores on the Ravens in particular. Although thinking aloud did not affect younger adult performance on any tasks, or older adult performance on three of four tasks, think-aloud older adults performed better than silent controls on the Ravens in two experiments. The authors were unable to account for the effect, but the use of a procedure that prompted regularly for continued verbalization combined with prolonged solution times of think-aloud older adults (there were no time limits) suggest that perceived pressure to provide useful information may have discouraged guessing or reduced older adult decisiveness of initial response selections.[9] In

[9] One interesting revelation of Study 2 think-aloud data that I do not report in the results is that incorrect responses are frequently caused by neglecting to consider alternative responses after identifying one that is compatible with an accurate inference. The data suggest that it is possible, at least in the absence of time limits, for many participants to

the present investigation, prompting is minimized to reduce the likelihood of reactivity. It is noteworthy however that whatever improved think-aloud older adult scores in Fox and Charness' (2010) experiments, if it is not eliminated by minimization of prompting, should tend to reduce rather than improve correspondence between findings and predictions in the present study.

Although think-aloud is an ontologically grounded method that should be expected to reveal veridical information under conditions of proper instruction, the data are incomplete (Ericsson & Fox, 2011), and are not specific enough to distinguish between functionally similar production theories. In most cases it would be unreasonable to expect verbal reports to be specific enough to be translated directly into actual productions (Singley & Anderson, 1989). However, this does not imply that verbalization cannot be used to test a more general process theory such as the present one, which treats functionally similar productions as a single theoretical category. At this grain-size, the question becomes whether or not mapping occurs, and information relevant to answering this question is available in the form of verbalized inferences about the objects present in correct item responses. As can be seen in Table 2 of Carpenter et al. (1990, p. 12), participants tend to verbalize properties of correct responses as they become available such as "ok, it should be a square." A verbalization specifying that a correct response must contain a specific object (such as a square) is an indirect indicator that mapping has occurred; namely, a missing object in the target analog (the bottom row or right column) has been identified. For example, a verbalization indicating that the answer to Figure 2 should contain a dot implies that productions have fired allowing for the mapping of objects that share a role at Level 2 of dissimilarity. Verbalizations of correct objects for rules at higher levels of dissimilarity are characteristic of the organizational knowledge that represents roles as unknowns.

Note that because verbal traces are incomplete (e.g., Ericsson and Fox, 2011), failure to verbalize does not necessarily imply that mapping did not occur. Given that it is the difference between cohorts that is of primary interest, this limitation is a concern only to the extent that groups differ systematically in the likelihood of emitting relevant verbalizations given the same basic cognitive state with respect to an item. There is no reason to assume that such differences exist. Nevertheless, because differences in verbalization frequency between persons add unwanted noise to the analysis, a multi-level generalized linear model (e.g., Breslow & Clayton,

---

raise their scores without acquiring any additional skill by simply taking the time to rule out multiple responses that are each compatible with a single inference.

1993) is used to model protocol data. This allows probability of critical verbalizations to vary as a random effect of individual participants and specific trials. The incorporation of random effects when observations are sampled from some larger hypothetical population is a safeguard against spurious conclusions that can arise from the tendency to overgeneralize fixed effects (Greenland, 2000). The model treats opportunities to verbalize a relevant object (one per rule) as "successes" or "non-successes" in a series of binomial trials, each corresponding to an individual rule within an item. This raises concerns about conditional dependence (Rosenbaum, 1984) that are considered in greater detail below.

An additional advantage of the multi-level model is that it makes it possible to rule out the possibility that predicted effects are driven by an alternative mechanism. It is possible that an observed interaction between cohort and dissimilarity of objects could be incidental to the fact that rules with dissimilar objects tend to appear in items with more rules. For example, it may be that individuals from both cohorts have equal baseline ability to map dissimilar objects (for example, when there is only one rule to infer), but that the capability of individuals from cohort 1 declines more than that of individuals from cohort 2 as the number of rules for an item increases. If this were the case, the predicted finding would not be very informative. By classifying each rule according to the number of total rules in the item from which it is sampled (e.g., all four rules for an item with four rules are classified as *4*), the model makes it possible to verify that an expected effect of dissimilarity is not a consequence of number of rules.

Other relevant verbalizations for testing the theory are available, but are unlikely to occur with the frequency necessary to test predictions against a binomial distribution. Previously-collected data reveal that participants also verbalize the conceptual opposites of correct objects, that is, verbalizations expressing the presence of an object in the correct response that is not actually contained within the correct response. Because there are more ways to execute a task incorrectly than correctly, the interpretation of these *incorrect objects* is not exactly the opposite of the interpretation of correct objects. The organizational knowledge for mapping dissimilar objects (the productions specified above) instantiates the rejection of roles and relations that do not permit coherent mapping of objects. This means verbalization of an incorrect object can have one of two basic interpretations, one of which pertains to the mapping process itself, and the other of which pertains to the higher level organizational knowledge that guides the process and is relevant to the theory. In the first case, a verbalization of an incorrect object could be

interpreted merely as evidence that a lower-level error occurred in the mapping process itself, in which case the verbalization is not immediately relevant to predictions. On the other hand, verbalization of an incorrect object could be interpreted as evidence that the object has been accepted even though coherent mapping was not achieved. In this case, an incorrect object implies an absence or failure of organizational knowledge, which, according to the theory, should be characteristic of individuals from earlier cohorts. Incorrect objects provide only limited information because it is impossible to rule out the former interpretation for any particular instance. Although relatively infrequent and ambiguous in reference, incorrect verbalizations are relevant as circumstantial evidence of failures to consider additional rules.

A final prospective category of verbalizations is roles, relations, and rules themselves. These verbalizations present obstacles to classification and analysis that are ultimately irresolvable, at least for present purposes. Like incorrect objects, roles, relations, and rules are relatively infrequent, but unlike correct or incorrect responses, they are not easily distinguished from other verbalizations. As Carpetner et al. (1990) note, participants tend to begin solving an item by passively encoding item properties. Thus, a prospective role (e.g., "they all have a square") may be verbalized whether the participant has apprehended a possible role or not. Relations are less ambiguous (e.g., "it gets bigger"), but it is unclear whether they should be interpreted as relations, applicable to only one analog, or rules that are applicable to all analogs unless the participant surrenders this information spontaneously. Finally, and most importantly, there is a crucial theoretical distinction between correct and incorrect objects on the one hand, and rules, roles and relations on the other as they relate to verbalization. Matrix reasoning items require applying a generalization to a particular instance. Thus, correct and incorrect objects will tend to be verbalized as they are initially represented by participants (as objects rather than roles) regardless of whether the objects are similar or dissimilar. In contrast, probability of verbalizing roles and relations is likely to be systematically related to dissimilarity of objects because verbalizations are generated more readily when familiar referents are available (Ericsson & Simon, 1993). This means rules with similar objects are more likely to elicit verbalizations of rules, roles, and relations than rules with dissimilar objects. The reason is that roles and relations are synonymous (or nearly synonymous) with objects and analogs when objects are similar; that is, parts of the matrix can be verbalized as they appear. In contrast, roles and relations must be verbalized as abstract referents when objects and analogs are dissimilar unless the participant

translates roles into objects while verbalizing (there is no reason to assume that participants represent complex roles and relations in form of words and phrases just because I have presented them that way). Because of this confound, roles, relations, and rules are not considered in the following analyses. Although this does limit the resolution of conclusions relative to what could be inferred with more precise information, objects alone are sufficient for testing the prediction that differences in mapping of objects cause differences in performance between cohorts.

Predictions

Predictions are relatively straightforward. Overall scores of cohort 2 should exceed those of cohort 1 if the samples are representative. If differences in the ability to map dissimilar objects are responsible for differences in scores between cohorts, level of dissimilarity should interact with cohort such that cohort 2 participants are more likely to verbalize correct objects, particularly at higher levels of dissimilarity. Assuming sufficient frequencies of incorrect objects, cohort 1 participants should verbalize these objects more frequently, particularly when objects of a rule are dissimilar. Number of rules in an item is confounded with dissimilarity of objects within its rules, but it is expected that the predicted interaction between cohort and dissimilarity is not dependent on number of rules.

Method

Participants. Forty older (mean age: 71 years, 26 women) and 40 younger (mean age: 19 years, 24 women) adults participated in the study. Older adults (cohort 1) were recruited from a community sample and were paid 10 dollars per hour to participate. Younger Adults (cohort 2) were recruited from the undergraduate participant pool and received course credit for participating. Most cohort 1 participants were born around the same time as participants comprising the cohort 1 data analyzed in Study 1. Like their counterparts from Study 1, cohort 1 participants of Study 2 are likely to be psychometrically superior, on average, to cohort 2 participants relative to cohort. Cohort 1 participants had an average of about 17 years of education, while the vast majority of cohort 2 participants were in their first year of college at the time of testing.

Materials. All 36 test items from Set-II of the Raven's Advanced Progressive Matrices (Raven, 1965) were used. In the ascending order condition, all items were presented in their original order. In the non-ascending condition the items were presented in a single pseudo-

random order that maximizes homogeneity of difficulty based on the pass rates reported in Forbes (1964). In both versions participants were asked to make confidence judgments after each item for a related study. Mitchum and Kelley's (2010) procedure was used, where participants are asked to type in a percentage value between 12 (chance performance) and 100 (certainty) reflecting their estimates of the probability that the item was answered correctly. Both versions of the test were presented by computer on a 19-inch monitor.

Procedure. Participants were tested individually. Those in the think-aloud condition completed a warm-up procedure (see supplemental materials of Fox et al., 2011) consisting of several exercises designed to teach participants to think aloud and discourage them from introspecting or explaining their thought processes. Participants in the silent group did not complete the seven-minute procedure. Participants were situated at the computer to complete the Ravens. The experimenter read the instructions aloud and requested that participants in the think-aloud condition begin verbalizing as they solved the four practice items. Because concurrent verbalization tends to prolong solution times (Fox, et al., 2011), no time limit was imposed to ensure that process data from participants of varying ability were available for every item. Participants were prompted to continue verbalizing very infrequently. Experimenters were instructed to allow participants to remain silent until the experimenter believed the participant had forgotten the instruction to think aloud or until commencement of a new item. The experimenters were apprised that participants should not feel compelled to verbalize if they do not have anything to say.

Encoding of Verbal Protocols. The same item classifications from Study 1 were used along with additional materials. Because there are multiple ways of defining correct and incorrect objects, it was necessary to develop specific criterion rules for classifying verbalizations. For the most part, these rules parallel Carpenter et al.'s (1990) rules, but were conceived so as to maximize compatibility with verbalizations. As Table 3 shows, it was necessary in 2 cases (items 4 and 10) to consolidate two of Carpenter et al.'s (1990) rules into a single more general criterion rule. In these cases a verbalization was counted if it applied to the consolidated rule or if it applied to only one or both of the initial rules.[10] For example, item 10 has two rules in Carpenter et al.'s (1990) taxonomy, one involving length of lines, and one

[10] Participants who verbalized an object for initial one of the initial rules tended to verbalize the other as well. Consolidation helps to address the problem of logical dependency.

Table 3

Item and Rule Classifications for Studies 1 and 2

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
1	Distribution-of-three-values	2	Shape[e]	Diamond	2
			Orientation of lines	Right	2
	Constant-in-a-row	1	Number of lines	Three	1
2[a]	Quantitative-pairwise-progression	1	Placement of bars	Middle	1
	Constant-in-a-row	1	Number of segments[e]	Three	1
3	Quantitative-pairwise-progression	1	Positions of objects[e]	Aligned	1
	Constant-in-a-row	1	Shape of objects	Hexagon	1
4	Quantitative-pairwise-progression/constant-in-a-row	2	Shape or constitution of object[e]	Square or dot	1
5	Quantitative-pairwise-progression	1	Amount of black[e]	One unit	1
	Constant-in-a-row	1	Shape	"L"	1
6	Quantitative-pairwise-progression	1	Total number of dots (horizontal)[e]	Three	1
	Constant-in-a-row	1	Orientation of dots	One line/one high	1

Table 3, continued

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
7	Figure-addition or subtraction	1	Presence of objects[d]	Cross	2
8	Distribution-of-three-values	1	Horizontal shading	Slanted lines	2
		1	Vertical shading	Black	2
9	Figure-addition or subtraction, constant-in-a-row	2	Shape or constitution of object[e]	Square	2
10	Quantitative-pairwise-progression/constant-in-a-row	2	Shape or constitution of object[e]	Square	1
11[a]	Figure-addition or subtraction	1	Presence of objects[d]	Vertical lines	2
12	Figure-addition or subtraction	1	Presence of object[e]	Dot	2
13	Distribution-of-three-values	2	Shape[e]	Square	2
			Orientation of lines	Left	1
	Constant-in-a-row	1	Number of lines	Three	1
14	Quantitative-pairwise-progression	1	Position of dot	Top	1
	Constant-in-a-row	1	Position of circle[e]	Bottom	1
15[a]	Quantitative-pairwise-progression	1	Background	Wavy	2
	Figure-addition or subtraction	1	Shape[e]	Square	2

Table 3, continued

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
16	Figure-addition or subtraction	1	Presence of object[e]	Circle	2
17	Distribution-of-three-values	1	Properties of line (curvy versus straight/dotted versus solid)[e]	Straight/dotted	2
	Constant-in-a-row	1	Configuration of lines	Triangle	1
18[ab]	NA	2	Type of line[e]	Curved	2
			Orientation	Slanted	2
19[ab]	NA	2	Presence of objects[d]	Circle	2
20[a]	Figure-addition or subtraction	1	Presence of objects[d]	Bars	2
21[a]	Quantitative-pairwise-progression	2	Outer shading	Black	2
			Basic shape[e]	Hourglass	2
	Constant-in-a-row	2	Middle shading	Black	2
			Orientation of object	Vertical	2
22	Distribution-of-two-values	3	Presence of circle	Present	3
			Presence of square[e]	Present	3
23	Distribution-of-two-values	4	Presence of circle[e]	Present	3
			Presence of cross	Present	3

Table 3, continued

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
			Presence of dot	Present	3
24[a]	Quantitative-pairwise-progression	1	Placement of horizontal lines[e]	Bottom	1
	Constant-in-a-row	1	Placement of vertical lines	Right	1
25[a]	Quantitative-pairwise-progression	1	Proportion of square with stripes	All	1
	Constant-in-a-row	3	Position of shading in circle[e]	Right side	3
26	Quantitative-pairwise-progression	1	Orientation of curve	Vertical	3
	Distribution-of-three-values	1	Presence of thick line[e]	Present	3
27	Distribution-of-three-values	2	Basic shape[e]	Circle	3
			Elongation/presence of fold	Elongated/folded	3
28[a]	Distribution-of-three-values	4	Type of horizontal lines	Straight	2
			Type of vertical lines[e]	Double squiggly	2
29	Distribution-of-three-values	3	Relative lengths of lines[d]	Short in the middle	2
			Orientation	Open on the right	2
			Basic shape[e]	"C"	3

Table 3, continued

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
30[a]	Quantitative-pairwise-progression/distribution-of-two-values	3	Presence of dotted shading	Present	3
			Presence of horizontal lines	Present	3
			Presence of vertical lines[e]	Present	3
31	Distribution-of-three-values	1	Presence of slanted lines	Present	2
	Distribution-of-two-values	3	Presence of blank box[e]	Present	3
			Presence of vertical lines	Present	3
32	Distribution-of-three-values	1	Presence of straight lines	Present	1
	Distribution-of-two-values	3	Presence of blank slice[e]	Present	3
			Presence of semicircle	Present	3
33	Figure-addition or subtraction	2	Number of black dots[e]	One	2
			Number of white dots	Two	2
34	Distribution-of-three-values	3	Basic shape[e]	Bracket	3
			Number of dots	Two	2
			Relative lengths of lines[d]	Long in the middle	2
	Constant-in-a-row	1	Shape of lines	Wavy	2

Table 3, continued

Item number	Carpenter et al.'s (1990) rules	Number of rules	Criterion rules for Study 2	Example of correct object	Level of dissimilarity
35	Distribution-of-two-values	3	Presence of dotted line (top)[d]	Present	3
			Presence of dotted line (bottom)[d]	Present	3
	Constant-in-a-row	1	NA		
36	Distribution-of-two-values	4	NA		
	Constant-in-a-row	1	NA		

Note. Any two of Carpenter et al.'s (1990) rules separated by a slash both apply to the same criterion rule or were combined into a single criterion rule that is more compatible with verbalizations (and vice versa). Carpenter et al.'s (1990) rules that correspond to absent objects are not presented (items 22, 23, 31, 32, 34, 35, and 36). Verbalizations of absence were relatively infrequent and were not included in analyses.

[a] Carpenter et al. (1990) do not report their own classification of item; [b] item cannot be classified based on Carpenter et al.'s (1990) taxonomy (p. 431); [c] criterion rules are not identical to Carpenter et al.'s (1990) rules; [d] yielded no relevant verbalizations for Study 2; [e] used in primary analysis for Study 2.

involving the distance of these lines from one another. The result is an apparently holistic transformation of a small square into a large square. Thus, participants were more likely to emit verbalizations relevant to the two rules combined, including phrases such as "gets bigger." The two original level 1 rules were combined into a single level 1 rule, but verbalizations relevant to only one of the two original rules (length of lines or distance between lines) were counted as verbalization of the consolidated rule. Figure-addition or subtraction rules that call for more than one object in a correct response also present a practical problem because in these cases verbalization of a single object is not necessarily a sufficient indicator that mapping has been accomplished. The correct responses to items 7, 11, 19, and 20 in particular include almost every object, which compromises the interpretation of verbalizations that refer to only one or two correct objects. The tendency of some participants to conjoin several correct objects into single referents, although expected, introduces more ambiguity. This problem was addressed by excluding these items from formal analyses. The cost of this omission is minor because the five figure-addition or subtraction items with only one relevant object per rule are preserved (items 9, 12, 15, 16, 32). The final 65 criterion rules are shown in Table 3.

To qualify as a correct object, a verbalization had to refer to the object (or some phrase that implicates the object) and contain either terms or phrases that suggest hypothetical thinking, such as "should," "probably," "will have," "alright, so…" and so on, or context had to suggest that the participant was not merely verbalizing properties of the items. Consistent with Carpenter et al.'s (1990) analysis, protocols collected prior to the study revealed that participants tend to begin solving an item by encoding objects passively prior to verbalizing objects contained in the item response. The cycle is repeated as additional inferences are made. Participants are more likely to remain silent during the encoding phase; that is, participants who verbalize very little are more likely to verbalize properties of their answer choice than to describe features of the matrix. For example, it is common for a participant to remain silent for most of the time spent on an item, and then verbalize an object immediately prior to selecting a response containing that object. For this reason, it was decided that isolated verbalizations occurring immediately prior to response selections were to be counted as correct or incorrect objects because this is a point at which objects that participants believe belong in the answer should be active in working memory. Relevant verbalizations were classified as correct objects irrespective of whether the referenced object is contained in incorrect responses. The only exception was objects contained within all

eight response choices, which were not classified. The following are some actual examples of correct object verbalizations from the protocols, and additional examples are provided in Appendix A:

Item 3: "I need three squares across."

Item 12: "Should keep a dot."

Item 17: "It's actually going to be a full triangle on the end."

Item 23: "It would be a square."

Incorrect objects are identical to correct objects except that they are *not* part of the correct answer. Encoding of incorrect objects required a more conservative judgment because it is not possible use accurate solutions as a reference point for top-down inferences about the meaning of a verbalization. Verbalizations of incorrect objects must be explicit enough to rule out the possibility that a participant has adopted an idiosyncratic representation that is foreign to the coder but nonetheless consistent with a correct response.

It is possible in theory for verbalization of the numbers of response choices to reveal information that is practically identical to the information revealed by verbalization of a correct object or incorrect object. Participants commonly narrow their choices down to two or three responses, which they may identify by number. It was decided that, although frequent and potentially relevant, these verbalizations would not be included in analyses for two related reasons. First, inclusion of answer choices would undermine the logical independence of verbalizations and answer choices. Second, participants are not necessarily logically consistent in the numbers they identify as possible answers. A participant may verbalize that the answer is either response 1, 2, or 3 even though these three responses have no object in common. Finally, is also possible for participants to change which responses they have ruled out without verbalizing information relevant to the cause of the change. The potential for invalid data would increase over the course of time needed to solve an item as participants continually update eliminated responses. Because Ericsson and Simon's (1980) theory does not assume that participants are incapable of making deductive errors or that they have infallible working memory, verbalized numbers representing response choices are not considered admissible as data.

The author encoded the verbalizations three times over a period of about three months in an effort to prevent changes in criterion from affecting the results. During this process, cohort of

participants was concealed, although it was often possible to identify cohort by verbalizations. Similarly, protocols were also encoded item by item rather than participant by participant to minimize expectations that could arise from becoming familiar with personal attributes of participants. The three sets of 1,440 item encodings apiece (36 items times 40 participants) were compared to establish which items changed from one set to another in terms of the number or composition of correct objects. In general, encodings of all three sets were very similar, the most common discrepancies being inclusions or exclusions of verbalizations that are interpretable as correct objects only in the context of surrounding verbalizations. Because differences were almost always reflected by a change in the number of correct objects, Spearman correlations between the number of correct objects per participant per item are a suitable indicator of the level of reliability. The three correlations (1 and 2, 1 and 3, and 2 and 3) each exceeded $r = .9$. In most cases, the third series of encodings was used as data.

Multilevel Model for Verbalizations. The probability of emitting a relevant verbalization of a given kind for a given rule can vary as a function of individual participant and item properties for reasons that are irrelevant to predictions, increasing noise and even the opportunity for false discovery. To minimize the influence of these effects on results, and maximize the generalizability of results, a multi-level generalized linear model was used to model responses at the level of individual rules and at the level of participants. The model is a simple linear growth model (Raudenbush & Bryk, 2002) that is adapted with a logit link function to accommodate a binomial distribution of responses. A hypothetical true probability of verbalizing for a specific rule under specific conditions is replaced with multiple individual probabilities that are normally distributed about a mean of various random effects. All variables were dummy coded to have a meaningful value of zero such that intercepts represent the level of 1 for each variable (0 = level 1 dissimilarity, 1 = level 2 dissimilarity, 2 = level 3 dissimilarity; 0 = one rule, 1 = two rules, 2 = rules; and 0 = cohort 1, 1 = cohort 2). Data were analyzed using HLM software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011).

Because verbalizing versus not verbalizing is a dichotomous variable that is unlikely to be normally distributed, a link-function is used to transform the binomial distribution of "successful" and "unsuccessful" verbalizations to the more familiar regression-like structural model at the level of rules. If Y_{ti} is the number of successful verbalizations emitted in m_{ti} opportunities, then

$$Y_{ti} \mid \phi_{ti} \sim B(m_{ti}, \phi_{ti}),$$

expresses the binomial distribution of Y_{ti} for m_{ti} trials with a predicted probability of ϕ_{ti}. The logit link-function transforms this distribution to the rule-level structural model. Probability of success is the log-odds, η_{ti}, where

$$\eta_{ti} = \log\left[\frac{\phi_{ti}}{1 - \phi_{ti}}\right] = \pi_{0i} + \pi_{1i}(\text{dissimilarity})_{ti} + \pi_{2i}(\text{number of rules})_{ti} + e_{ti},$$

or the probability that person *i* emits a relevant verbalization at trial *t* as a function of person *i*'s probability of verbalizing a correct object for a rule with level 1 dissimilarity that belongs to an item with one rule, π_{0i}, and the change in this probability attributable to dissimilarity of objects, π_{1i}, number of rules, π_{2i}, and rule-specific error, e_{ti}. The participant-level model assumes the form,

$$\begin{aligned}
\pi_{0i} &= \beta_{00} + \beta_{01}(\text{cohort})_i + r_{0i} \\
\pi_{1i} &= \beta_{10} + \beta_{11}(\text{cohort})_i + r_{1i} \\
\pi_{2i} &= \beta_{20} + \beta_{21}(\text{cohort})_i + r_{2i},
\end{aligned}$$

where β_{00} is the mean probability of verbalizing a correct object for a rule with level 1 dissimilarity that belongs to an item with one rule, β_{01i} is the change to this probability associated with person *i*'s cohort, and r_{0i} is the unique effect associated with person *i*. Note that the random effects components imply that the model allows participants to have unique probabilities of verbalizing, but also unique *changes* in probabilities of verbalizing, even though the latter is not expected to be necessary. For the sake of simplicity, information sought from the intercepts and slopes of the three equations can be translated to the following six questions:

1. On average, are participants' probabilities of verbalizing different from zero for rules with level 1 dissimilarity that belong to an item with one rule?
2. On average, is the probability of verbalizing greater for participants in one cohort than another?
3. Does dissimilarity of objects affect the average probability that participants verbalize?
4. As dissimilarity of objects changes, is the change in probability of verbalizing greater on average for one cohort than the other?
5. On average, do participants' probabilities of verbalizing correct objects depend on the total number of rules in the item that a particular rule belongs to?

6. As the total number of rules in an item increases, is the change in probability of verbalizing a correct object greater on average for one cohort than another?

Note that only question 4 is of principal relevance to predictions, although 6 may be of general interest and serves the function of ruling out a potential answer to 4 that is irrelevant to predictions.

A meaningful comparison of observations to a frequency distribution requires conditional independence of observations; that is, a success or failure on one trial should not influence the probability of success on a subsequent trial (Rosenbaum, 1984). The present data present a threat to conditional independence because there are multiple observations per item. Reasonable arguments could be made that successfully mapping objects for one rule of an item increases the likelihood of mapping objects for a second rule of the same item. Statistical treatment of the problem (which requires assumptions in addition to the considerable number that are already taken for granted by multilevel models) is less ideal than a simple demonstration that potential violations of independence do not imperil the validity of conclusions. A satisfactory demonstration would, if possible, verify that the most conservative analysis reveals the same results as an analysis that ignores the concern entirely. To accomplish this goal, I present an analysis that minimizes the likelihood of conditional dependence by including only one rule per item. The simple selection procedure selects the rule that (1) has the highest level of dissimilarity, and in case of a tie, (2) has the most successes or verbalizations. Both criteria are non-arbitrary, the first because there are relatively few level 3 rules, and the second because comparisons to binomial distributions yield the most accurate estimates when probabilities are not too close to zero or one (e.g., Carriere, 2001). Probabilities are considerably lower than 50 percent when considering all 65 rules. Despite this omission, it is noteworthy that an identical analysis that includes all 65 rules yields results very similar to those reported below, including the effect size of the slope for question 4 above.

Results

Performance. Means and standard deviations of scores and solution times for every condition are presented in Table 4. Having provided these descriptives, I report only simple comparisons that are relevant to ruling out confounds for forthcoming analyses. The overall difference in scores between cohort 1 (M = 11.20, SD = 5.29, 95% CI [9.56, 12.80]) and cohort 2 (M = 16.30, SD = 5.46, 95% CI [14.60, 18.00]) is roughly one standard deviation (d = .95),

which is somewhat smaller than the typical Flynn effect of the Ravens for two cohorts separated by about fifty years (about d = 1.5; Flynn, 2007). In fact, cohort 2 scores in the present study are lower than the scores of cohort 2 participants from Study 1 which, were around 60 percent (about 21 correct responses) in each constituent sample. In contrast, cohort 1 scores appear to fall within the range that is typically observed in the contemporary cognitive aging literature. For example, cohort 1 participants performed a little better than Babcock's (2002) 60 to 90 year-olds (n = 282), about 43-percent of whom solved more than 11 items correctly (means were not reported) under the constraint of a 20-minute time limit. Table 5 shows a comparison of Study 1 and Study 2 pass rates. Whatever the cause, lower-than-expected scores among cohort 2 participants do not present any serious threat to the interpretation of the analyses that follow given that obtained scores fall well within the normal range of contemporary young adults. As Table 4 shows, there are no effects of order as scores were similar in both the ascending (M = 14.00, SD = 4.62, 95% CI [12.60, 15.40]) and random conditions (M = 13.50, SD = 6.13, 95% CI [11.60, 15.40]; d = .09).

A minimum necessary condition for the validity of process data is the non-reactivity of think-aloud verbalization with respect to scores. There was no sign of a main effect of verbalization as mean scores of think-aloud participants (M = 13.20, SD = 5.15, 95% CI [11.60, 14.80]) were comparable to silent participants (M = 14.30, SD = 5.60, 95% CI [12.60, 16.00]; d = .20). As Table 4 shows, there is almost no trace of a verbalization effect among participants in cohort 1, regardless of whether they completed the test in ascending or random order. It is unclear whether this non-replication of Fox and Charness' (2010) is the result of using a procedure that minimized prompts to continue verbalizing. Within cohort 2, the think-aloud group (M = 15.50, SD = 4.85, 95% CI [13.40, 17.60]) scored somewhat lower than the silent group (M = 17.10, SD = 6.08, 95% CI [14.40, 19.80]), but the effect size falls well within the range of expected error (d = .29). Overall, no evidence was found that verbalization altered overall performance of participants in either cohort.

Table 4

Mean Scores and Solution Times for Study 2

Condition	Cohort 1 (n = 40)	Cohort 2 (n = 40)	Total (N = 80)
Score	11.20 (5.29)	16.30 (5.46)	13.70 (5.37)
Think-aloud	11.00 (5.45)	15.50 (4.85)	13.20 (5.15)
Ascending	11.10 (5.24)	16.30 (3.95)	13.70 (4.60)
Random	10.90 (5.65)	14.60 (5.74)	12.80 (5.70)
Silent	11.40 (5.13)	17.10 (6.08)	14.30 (5.60)
Ascending	11.00 (4.14)	17.50 (5.14)	14.30 (4.64)
Random	11.80 (6.11)	16.70 (7.01)	14.30 (6.56)
Solution time	38.70 (20.50)	21.80 (8.45)	30.20 (14.50)
Think-aloud	38.30 (18.10)	25.20 (10.80)	31.80 (14.50)
Ascending	34.90 (15.20)	26.50 (11.90)	30.70 (13.60)
Random	41.70 (20.90)	24.00 (9.76)	32.90 (15.40)
Silent	39.10 (22.80)	18.30 (6.05)	28.70 (14.40)
Ascending	42.40 (32.20)	18.10 (5.86)	30.30 (19.10)
Random	35.80 (13.40)	18.40 (6.24)	27.10 (9.82)

Note. Scores are mean number of correct responses out of 36, and solution times are mean number of minutes from commencement of the first item to completion of the last. Standard deviations are presented in parentheses.

Solution Time. Overall, participants from cohort 1 (M = 38.70) spent considerably longer solving the 36 items than participants from cohort 2 (M = 21.80), but showed far greater variability as indicated by substantial disparity in standard deviations between cohorts (cohort 1: SD = 20.50, cohort 2: SD = 8.45). There was no observable effect of order as participants in ascending (M = 30.50, SD = 16.30, 95% CI [25.40, 35.64]) and random (M = 30.00, SD = 12.60, 95% CI [26.10, 33.90]) conditions spent about the same amount of time solving the items (d = .03). Although verbalization is known to prolong solution times, think-aloud participants (M = 31.80, SD = 14.50, 95% CI [27.30, 36.30]) did not spend considerably longer on average than

silent participants (M = 28.70, SD = 14.40, 95% CI [24.20, 33.20]), although the trend is in the expected direction (d = .21).

Table 5

A Comparison of Mean Pass Rates from Studies 1 and 2

Study	Cohort 1	Cohort 2	Effect size (d)
1	.60 (.29)	.59 (.26)	–.04
2	.32 (.19)	.45 (.25)	.59

Note. Mean pass rates are the average rate of correct responses across all 36 items.

Verbal Protocols. As expected, verbalizations of correct objects were relatively frequent. Two rules that did not elicit any correct objects are excluded from the analyses that follow (one in item 29 and one in item 34). These rules were not necessarily more challenging, but refer to objects that either do not lend themselves to verbalization or elicit verbalizations that are too ambiguous to be encoded. In particular, both refer to the relative lengths of the three lines that constitute a shape. The possible roles of *equal length, long in the middle,* or *long on the ends* are not easily translated into simple objects for verbalization. The final set, a series of 2,600 trials representing forty participants (65 trials each), constitutes one of the largest analyses of think-aloud data ever reported in a single study.

Parameter estimation was achieved using penalized quasi-likelihood (Breslow & Clayton, 1993), which yielded results that are nearly indistinguishable from the full maximum-likelihood procedure of adaptive Gaussian quadrature. In the following analyses I report r effect sizes and confidence intervals calculated from coefficient estimates and their standard errors to improve interpretability and translatability of results. The coefficients themselves are presented with both standard errors and effect sizes in Table 6. As noted above, potential violations of conditional independence are addressed by presenting an analysis that consists of one rule from each of the

Table 6

Verbalization of Correct Objects as a Function of Cohort

	Fixed effects					
	Intercept			Slope (cohort)		
Effect	Estimate	*r*	95% CI	Estimate	*r*	95% CI
Independent trials (n = 31)						
Intercept, β_{00}	–.41 (.20)	–.31	[–.57, .00]	–.08 (.33)	–.04	[–.35, .27]
Level of dissimilarity, β_{10}	–.22 (.13)	–.26	[–.53, .06]	.38 (.17)	.33	[.02, .58]
Number of rules, β_{20}	–.21 (.13)	–.25	[–.52, .07]	.01 (.16)	–.01	[–.32, .30]
Dependent trials (n = 65)						
Intercept, β_{00}	–.67 (.28)	–.35	[–.60, –.04]	.01 (.39)	.00	[–.31, .31]
Level of dissimilarity, β_{10}	–.05 (.08)	–.10	[–.40, .22]	.20 (.11)	.27	[–.05, .54]
Number of rules, β_{20}	–.23 (.08)	–.41	[–.64, –.11]	.04 (.11)	–.06	[–.36, .26]

	Random effects			
Effect	Variance component	*df*	X^2	*p*
Independent trials (n = 31)				
Intercept, r_{0i}	1.01	38	93.10	<.001
Level of dissimilarity, r_{1i}	0.07	38	44.10	.23
Number of rules, r_{2i}	0.11	38	55.60	.03
Rule-level error, e_{ti}	0.91	38		
Independent trials (n = 65)				
Intercept, r_{0i}	1.19	38	130.00	<.001
Level of dissimilarity, r_{1i}	0.02	38	42.10	.30
Number of rules, r_{2i}	0.02	38	44.01	.23
Rule-level error, e_{ti}	0.94	38		

Note. Estimates are the log-odds of verbalizing a correct object and correspond to the level of *1* for each variable (level 1 dissimilarity, one rule, and cohort 1). Standard errors appear in

parentheses.

31 items that elicited verbalizations of correct objects. Following this analysis, I briefly discuss the homologous findings of the full set of 65 rules.

The cohort 1 probability of verbalizing a correct object for a rule with level 1 dissimilarity that belonged to an item with one rule was about 40 percent (r = –.31, 95% CI [–.57, .00]), but neither cohort exhibited a higher probability of verbalizing than the other (r = –.04, 95% CI [–.35, .27]). The intercept for dissimilarity reveals that participants were, on average, more likely to verbalize a correct object for similar rules (r = –.26, 95% CI [–.53, .06]). As predicted, cohort 1 and cohort 2 had differing slopes for dissimilarity as cohort 1 verbalization of correct objects declined more than cohort 2 verbalization as rules became more dissimilar (r = .33, 95% CI [.02, .58]). Figure 7 shows the modeled change in probability of verbalizing for every participant as a function of dissimilarity and cohort.

Although the predicted finding was obtained, it is possible that the higher cohort 2 probability of verbalizing correct objects at higher levels of dissimilarity is incidental to the fact that rules with dissimilar objects tend to appear in items with more rules. The intercept revealed that the number of rules contained by an item from which a rule was sampled did not affect probability of verbalizing a correct object (r = –.25, 95% CI [–.52, .07]). Contrary to the hypothesis above, probability of verbalizing as a function of proportion of rules was unaffected by cohort (r = –.01, 95% CI [–.32, .30]). That is, the relative slopes of dissimilarity for the two cohorts (slopes in relation to one another) were unchanged by proportion of rules. The finding can be conceptualized by considering slopes of the lines in Figure 7, which are presented at the mean number of rules (M = 2.58). The observed effect size of r = –.25 implies that the slopes decrease somewhat for participants in either cohort as number of rules increases (i.e., the lines rotate clockwise). However, the mean slopes of the two cohorts do not change in relation to one another. According to the model, participants in neither cohort were impaired more than participants in the other by an increase in the number of rules in an item across levels of dissimilarity. This null finding lends support to the theory by effectively showing that the predicted finding is not dependent on number of rules, but it is important to consider that the properties of Ravens items are not fully counterbalanced, which restricts the range of

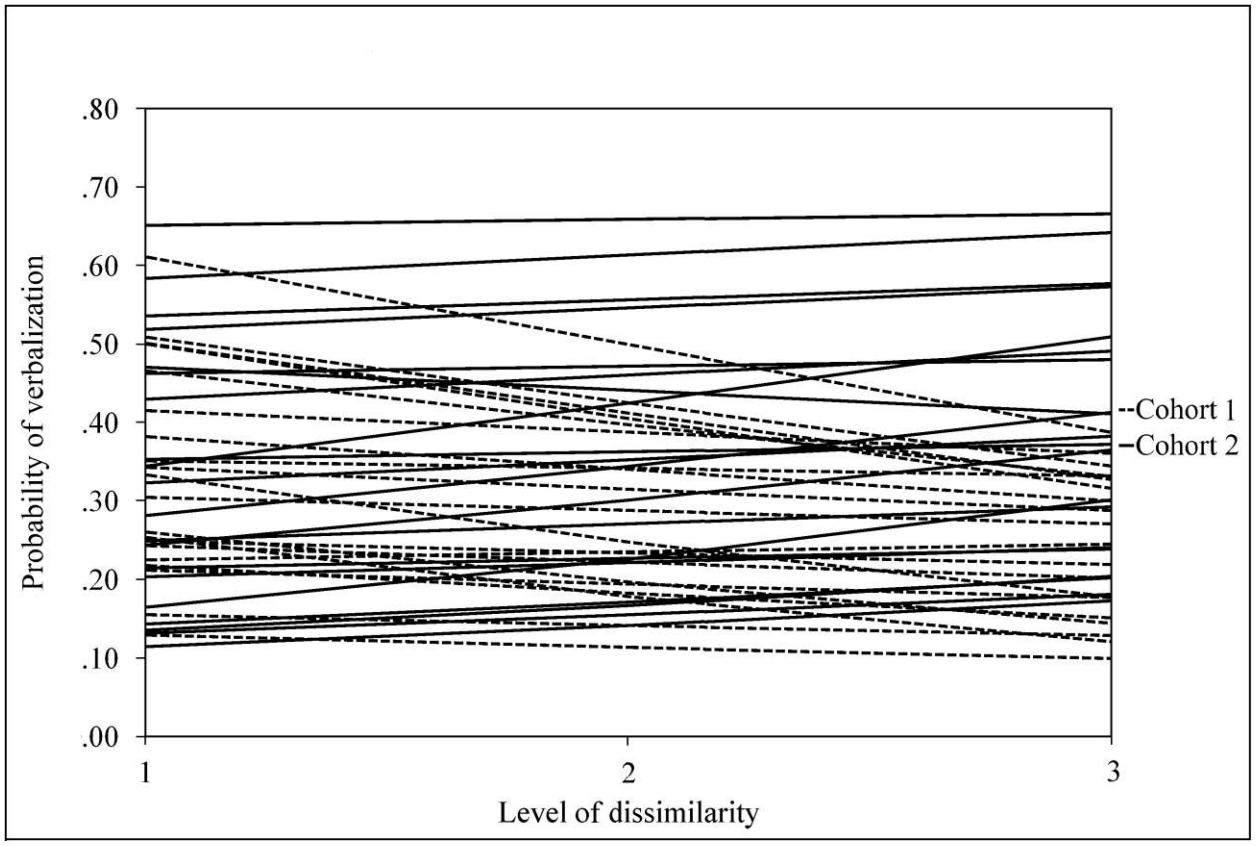

Figure 7. Modeled change in probability of verbalizing a correct object as a function of level of dissimilarity and cohort. The graph depicts the slopes of the forty think-aloud participants held constant at the mean number of rules in an item (M = 2.58).

observations. For example, there are no items with only one rule that are classified as level 3. Thus, it is possible that a difference in slope between cohorts (or age groups) for dissimilarity would be found in a set of fully counterbalanced items. However, such a finding would not constitute an alternative explanation of the present finding, which shows that the predicted disproportionate impairment of cohort 1 performance as a function of dissimilar objects does not depend on number of rules.

The corresponding statistics for all 2,600 trials of all 65 rules are also presented in Table 6. I refrain from describing them in detail because the slopes are very similar to those reported above. In particular, the slope representing the interaction between age and dissimilarity is similar. However, there is an apparent difference in main effects between the two analyses, as the larger set reveals a greater effect size for number of rules. It is important to consider that rules were selected for inclusion in the smaller set of 31 on the basis of verbalization frequency and level of dissimilarity. The smaller set contains only the 31 rules that elicited the most

verbalizations for their respective item. This would have tended to restrict the range of verbalization for comparing items with greater of fewer rules. Meanwhile, variation in verbalization frequency as a function of dissimilarity is similarly restricted in the larger set, which includes all of the rules that were excluded from the smaller set for eliciting relatively few verbalizations (some of which elicited very few verbalizations). In other words, although it is likely that the larger set violates conditional dependence, the apparent difference in main effects is most likely a methodological artifact of item properties and the incompleteness of verbal reports (Ericsson & Fox, 2011). However, because *changes* in verbalization frequency are relatively stable across participants, the interaction between dissimilarity and cohort manifests in either analysis. The random-effects variance components for dissimilarity, shown in Table 6, are relatively low, as between-subjects variation in slope does not exceed the limit of what would be expected if different participants are affected similarly by this variable (although there is evidence of heterogeneity in slope for number of rules). By no means does the absence or ambiguity of main effects contradict the large within-cohort effect sizes observed in Study 1 for dissimilarity and number of rules. An item with four rules can be much more difficult to solve than an item with one rule even if the probability that participants verbalize a correct object is higher for each of its four rules than the corresponding probability of verbalizing the one rule for the other item. There are no points awarded under normal testing conditions for correctly identifying three out of four correct objects.

Verbalization of incorrect objects occurred on less than one-percent of trials, ruling out the possibility of sophisticated analyses, but this frequency is adequate for simple analyses that treat verbalizations as count data. Because numbers of verbalizations are not normally distributed, non-parametric tests were used to compare mean ranks of Cohorts 1 and 2 by aggregating the relative placements of each participant in either group out of the overall total of 40 participants. Like the generalized linear model, this approach is ideal for comparing populations because it does not allow a finding to be driven or suppressed by a few extreme individuals. Overall frequencies of incorrect objects were low as verbalizations occurred on less than one-percent of trials at every level of dissimilarity (level 1: $M = .03$, level 2: $M = .06$, level 3: $M = .07$). In accordance with predictions, mean rank of cohort 1 was higher than mean rank of cohort 2 (cohort 1: $M = 24.70$, cohort 2: $M = 16.40$; $Z = -2.28$, $r = -.37$), as participants in cohort 1 were more likely to verbalize incorrect objects. There was no sign of an interaction, as differences

between cohorts in mean rank were comparable at each level of dissimilarity (level 1: $Z = -2.11$, $r = -.33$; level 2: $Z = -1.10$, $r = -.17$; level 3: $Z = -1.93$, $r = -.31$). Although participants in both cohorts verbalize incorrect objects relatively infrequently, cohort 1 participants verbalized them more often. The overall frequencies are too low for the absence of an interaction to have any theoretical significance. These results are compatible with the theory, but provide only circumstantial evidence that participants from cohort 1 are less likely to generate roles or relations that defy initial representations.

The protocols reveal additional information that is relevant to understanding the cognition of matrix reasoning and how it differs between cohorts. If the following observations seem superfluous, it is noteworthy that they are all replications of earlier observations made in the context of cognitive aging studies. These observations led to initial suspicions that abstract analogy is not quite an overlearned skill and that cross-sectional differences in scores are largely procedural differences in abstract analogical mapping.

It was found that participants of all ability levels were relatively unsystematic and prone to accepting incorrect objects. For example, a high-performing younger adult inferred correct objects for two rules (two dots and wavy lines) of one of the most difficult items, number 34:

> "The bottom row already has a one and a three which should be two dots."
>
> "Looks like it should be…curve wavy."

The participant almost failed to verbalize the correct overall shape (viz., a "C" or "bracket"):

> "Don't believe it should be bracket…but it might be some sort of bolder bracket; I don't know…I should go with that."

As this example shows, even relatively high performers occasionally accepted objects that, based on the mapping process, had not been established as correct (in this case, to the participant's advantage). Similar uncertainties or contradictions can be identified for virtually every participant as 39 out of 40 verbalized at least one incorrect object.

Several findings are compatible with the claim that task instructions and practice items are unlikely to impart a general task representation for individuals who map poorly. Frequent verbalization of a particular word by cohort 1 participants and low-performing cohort 2 participants is telling. The word, *pattern,* often embedded within the phrase *looking for a pattern,* was verbalized frequently by cohort 1 participants ($M = 12.50$), as many as 66 times by one participant. In contrast, cohort 2 participants verbalized this word infrequently ($M = 3.00$) with a

maximum of 13 verbalizations by a low performer. While it is true that *pattern* appears in the instructions, and an analogy is undoubtedly a complex pattern of sorts, ordinary use of the word connotes superficiality and would typically refer to something like a within-analog relation in a matrix reasoning item. Such a relation, by virtue of not referring to another analog, is necessarily superficial.

Perhaps most strikingly, several cohort 1 participants adopted a strategy of ruling out response choices by checking whether they are identical to figures in the matrix. The procedure was applied most frequently to more difficult items with dissimilar objects:

> *Item 23:* "We can eliminate one because that's a repeat of the right middle so I don't think we can use a repeat."
>
> *Item 27:* "Ooh, number eight is different; so is number seven; number five is different; alright, let's go to elimination here; number one can be eliminated; number two can be eliminated; four is out; what've we got left is number three, five, seven, and eight."

Such a strategy may seem irrational, but only if one takes for granted that the instructions and practice items confer a general task representation that applies to every item. If this assumption is true, it only applies to those individuals who map well enough to have the same basic problem representation regardless of whether an item's rules have similar or dissimilar objects. Participants who do not map well do not appear to perceive items containing rules with dissimilar objects as coherent analogs to what they learned in the instructions. For them, the goal of mapping similar objects is compatible with the putative goal of the task during the instructions, practice items, and some of the easier items, but is incompatible with this goal when items contain rules with dissimilar objects. In other words, more difficult items may prompt very low performers to adopt an alternative task representation that sustains the procedure of identifying similar objects, but in pursuit of an irrelevant non-analogical goal (e.g., "I don't see a pattern in this one"). Other cohort 1 participants expressed confusion as to the goal of the task despite performing well above chance, and at least one cohort 1 participant verbalized goal-related confusions repeatedly during the task after having solved all four of the practice items correctly.

Discussion

Results of Study 2 are consistent with predictions. Individuals born around 1990 scored higher than individuals born prior to 1950 on the test with the largest Flynn effect. Verbal protocols revealed that recent-born participants verbalized more objects compatible with correct mapping of dissimilar objects in particular, and tended to make fewer false inferences compatible with inaccurate mapping. No evidence was found that the interaction between cohort and dissimilarity was driven by number of rules in an item associated with an individual rule.

A methodological concern, given Fox and Charness' (2010) finding of higher scores among think-aloud older adults on the Ravens, was that thinking aloud may confer an advantage to older adults that undermines the validity of their verbal reports. However, there were no performance differences between think-aloud and silent conditions for either age group, suggesting that thinking aloud had little effect on the processes executed during the task. Whether non-reactivity would have also been observed with more frequent prompting as it was by Fox and Charness (2010) cannot be known, but for present purposes, there is no reason to assume non-validity. In fact, a post-study comparison of correct objects to the content of each of the eight response choices for every item revealed a correspondence rate of .96 across all 2,920 trials between verbalizations and responses. That is, 96 percent of responses selected by participants are compatible with verbalizations or absences of verbalizations based on encodings of the protocols. The corresponding rate when considering only successful trials is .84. Note that a mismatch between verbalization and item response does not imply that the cognitive state inferred from a verbalization did not occur at some point during the solution, which suggests that a rate of .84 for successful trials is likely to understate the correspondence of encoded verbalizations to correct mappings. In addition, a series of nonparametric comparisons revealed no systematic difference in rates of mismatch as a function of cohort and other variables. Overall, the results are compatible Ericsson and Fox's (2011) conclusion that think-aloud is a practicable and valid means of obtaining process-relevant data that is limited less by reactivity or validity than by completeness and resolution.

The *weakness* befitting a weak method (Anderson, 1987; Newell, 1973) is evident in the protocols, which revealed that even high performers, who frequently made successful inferences requiring abstract mappings, did so with little consistency, often changing inferences, and even recanting correct objects. Some participants in both cohorts, but especially low performers in cohort 1, exhibited behaviors consistent with incoherent task representations that appear to be the

result of attempting to map similar objects even when doing so defeats the purpose of the task. Regardless of cohort, the entire range of skill in the sample resembles the range of competency observed in the early stages of skill acquisition when procedures are not yet systematic enough to be rehearsed in any strict form (VanLehn, 1996). For reasons discussed earlier, there can be no strict algorithmic procedure for solving analogies, but no participants' protocols were consistent even with strict adherence to a procedure of mapping *as if to acknowledge* that there is no algorithmic procedure. Virtually every participant (n = 39) accepted at least one incorrect object.

No compelling evidence was found that that dissimilar objects are more difficult to map in general, and only tentative evidence was found that objects of rules belonging to items with more rules are more difficult to map. Even setting aside that individual rules belonging to items with many rules need not be more difficult, the absence of large main effects is uninformative. Anyone who is familiar with think-aloud reports would be unsurprised to learn that very easy level 1 rules do not necessarily elicit more verbalizations. For example, Carpenter et al. (1990) reported that constant-in-a row rules (level 1 in the present study) are verbalized less frequently than others. Participants vary substantially in their tendency to verbalize, and items vary substantially in their tendency to elicit verbalization, which tends to reduce the resolution of main effects, as can be seen by the large random-effects variance components in Table 6 for overall verbalization and rule-specific error. What is crucial is that *change* in verbalization as a function of dissimilarity is relatively stable from person to person, and that the predicted interaction between cohort and dissimilarity was observed in both analyses.

The major finding that the two cohorts have distinct slopes for dissimilarity of objects is both unambiguous, and compatible with Study 1. The effect size of r = .33 may seem small compared to the Flynn effect itself (about d = 1.5 or r = .60 for Ravens scores randomly sampled from two cohorts separated by 50 years) and Study 1 (r = .53). However, it is important to consider that the observed effect has a fairly large confidence interval and is not identical to test scores or pass rates of items. It is unclear exactly how large the effect size should be given that variation in successes and failures at executing single processes need not translate linearly to variation in correct responses if correct responses depend on the successful completion of multiple processes. Moreover, think-aloud participants in cohorts 1 and 2 differed in overall scores by less than one standard deviation (d = .87, r = .40).

It should not be forgotten that a difference between cohorts is, in this case, also a difference between age groups, but an age-related explanation of the results is not easy to defend. As far as I am aware, there are no theories in cognitive aging that would make the same rule-specific predictions as the current proposal. The claim that abstract rules tax older adults disproportionately because older adults have lower working memory capacity is subject to the same argument I presented when reviewing Carpenter et al. (1990). If individual differences in mapping dissimilar objects are caused by individual differences in procedural knowledge, then the conclusion that older adults possess less of this knowledge is more likely to be accurate than the conjunction that they possess less of this knowledge because of low working memory. While one might argue (contrary to Carpenter et al., 1990) that everyone possesses all of the necessary procedural knowledge, this alternative cannot explain why practice and training improve scores on tests so rapidly and effectively (e.g., Bors & Vigneau, 2001; Klauer & Phye, 2008). Moreover, this interpretation is incompatible with any straightforward interpretation of Study 1 findings, which confirmed predicted differences in pass rates between two groups of participants who were roughly the same age at the time of testing. A qualitative analysis of Study 2 protocols suggests that some cohort 1 participants do not even perform the same task as cohort 2 participants when items have dissimilar objects. Finally, and perhaps most decisively, the Flynn effect has to be caused by some cognitive mechanism that is distinct from the causes of age-related cognitive decline. As a knowledge-based theory, the present proposal is the only cognitive theory that satisfies the theoretical and empirical requirements of a cohort-related explanation of differences in scores.

Nevertheless, an age-related explanation cannot be dismissed entirely. As noted, Study 1 scores for both cohorts are much higher than Study 2 scores, especially for cohort 1 participants. By comparing two cohorts with identical overall pass rates to test a cognitive theory about cohort differences, Study 1 entailed an implicit assumption, namely, that raw scores on the Ravens test violate measurement invariance between cohorts with respect to specific cognitive causes. That predictions were confirmed suggests that a substantial proportion of cohort 1 participants who answered as many items correctly as cohort 2 participants solved a lower proportion of items with dissimilar objects, but just as importantly, a higher proportion of other items. From the vantage point of one large population, the assumption is that raw test scores either *overestimated* the ability of cohort 1 participants to map objects of a given level of dissimilarity relative to

cohort 2 participants, or *underestimated* the ability of cohort 1 participants to accomplish other essential problem solving goals relative to cohort 2 participants. Study 2 makes essentially the same assumption, and it is assumed that age-related mechanisms contributed to the greater disparity in scores. Whether or not the Ravens test violates measurement invariance in the manner predicted is a testable hypothesis.

CHAPTER SIX

GENERAL DISCUSSION

In general, investigators have tended to attribute differences in matrix reasoning to differences in working memory (e.g., Carpenter et al., 1990; Embretson, 1998; Primi, 2002). The present results are compatible with previous findings, but neither verify nor challenge claims about the role of working memory for reasons that I alluded to when reviewing Carpenter et al. (1990). What I have proposed cannot be mutually exclusive with (or complementary to) a psychological quantity unless that quantity refers to something that is logically distinguishable from procedural knowledge or observed performance. Such a quantity's (e.g., Daily et al., 2001) role as a source of individual differences can be understood only in light of how tasks are accomplished because differences in knowledge and procedure can be expected to change how much of such a quantity is needed to achieve a given level of observed performance (Ericsson & Kintsch, 1995). With that said, compelling evidence remains that differences in matrix reasoning are attributable, at least in part, to differences in the ability to preserve analogical rules (e.g., Embretson, 1998).

The findings challenge the seldom-acknowledged assumption that participants of various ability levels are all capable of generalizing task instructions to every matrix reasoning item. A rational analysis of analogical problem solving suggests that this assumption is problematic to begin with, but Study 2 provides tangible evidence that instructions and practice items do not necessarily prepare participants from earlier cohorts for the entire range of goals they must accomplish in order to achieve a high score. Cohort 1 participants were less likely to map dissimilar objects, and tended to commit errors that are characteristic of an incomplete task representation. In all likelihood, those individuals who possess the procedural knowledge needed to map dissimilar objects with any degree of consistency find that the goal of solving an item is more or less self-evident and constant, whereas those who cannot map these objects find the goal wayward and nebulous. To them, dissimilar objects defy their understanding of the instructions (e.g., "I don't see a pattern in this one").

Finally, the findings encourage cognitive aging researchers to be cognizant of the Flynn effect and its ramifications. The trend constitutes a major cross-sectional confound that is seldom mentioned in this literature (Zelinski & Kennison, 2007) even though its effect size is, for some

tests like the Ravens, larger than typical main effects observed in cross-sectional studies. In fact, Dickinson and Hiscock (2010) concluded after analyzing normative data from two versions of the Wechsler Adult Intelligence Scale (WAIS-R and WAIS-III) that cohort is responsible for the *majority* of the differences in cross-sectional scores obtained across subtests for groups separated by fifty years of age. In an earlier study, Hiscock (2007) estimated that only about one-third of the cross-sectional difference in Ravens scores is attributable to age. The present findings lend credence to concerns raised by others (Hofer & Sliwinski, 2001; Schaie, 2009) that effects of cohort and time period are understated by the most obvious interpretations of cross-sectional findings.

The remainder of this paper is devoted to considering the acquisition of weak-method procedural knowledge for mapping, identifying its ecological source, and proposing ways to test the theory. An important message throughout has been the indifference of between-subjects variation to psychological causes, and the misunderstandings that arise from interpreting variation as a causal entity. These misunderstandings are consequential when comparing distinct populations like those distinguished by a century of the Flynn effect. I conclude by showing why bottom-up (Borsboom et al., 2004; Cervone, 1997) theories like the present one are essential to making sense of such comparisons.

Analogy as a Weak Method: The Role of Example-Based Problem Solving

According to Singley and Anderson's (1989) theory, transfer is common productions, which implies that transfer can be the cause of rising scores on culture-free tests if there are productions that are common to affected tests on the one hand, and one or more real-world tasks or learning exercises on the other. If they exist, these tasks or exercises must be commoner today than a century ago, at least in regions that witnessed or are presently witnessing Flynn effects. According to what I have proposed thus far, the underlying productions must also pertain to the goal of mapping objects when relations between objects are not implied by objects themselves. One learning exercise in particular merits close consideration.

Contrary to prior assumptions, it has become increasingly evident in the last 25 or 30 years that participants learning to solve problems in domains such as math and science do not simply memorize verbal instructions and then apply them to problems (Anderson, 1987; Anderson & Fincham, 1994; Singley & Anderson, 1989). Instead, they rely on a process of *analogy,* mapping new problems to examples provided by instructions (e.g., Pirolli & Recker, 1994; Reder,

Charney, & Morgan, 1986; VanLehn, 1996). This realization was compelling enough to motivate Anderson and his colleagues, while upgrading the ACT* architecture into ACT-R, to augment declarative encoding of instructions with a process of analogy instantiated by productions. Instead of incorporating instructions into declarative knowledge as before, ACT-R incorporated *examples of problems solved by following these same instructions* (Anderson & Fincham, 1994, p. 1338):

> Although not going so far as to deny that other types of declarative knowledge might be sources for procedures, the emphasis has shifted to learning from examples. It is argued that initial use of these examples involves analogy and that production rules are compiled that summarize the analogy process. The major reason for this shift of emphasis to examples has been the research with acquisition of academic (mathematics, science, and computer programming) problem-solving skills and the evidence that subjects make heavy reference to examples in their initial attempts to solve problems in these domains.

Since then, a more formal analogical compilation process has been incorporated into the architecture (Anderson, 2007; Taatgen, 2003). It is now well-established that analogy serves a crucial function during the earliest moments of learning to solve new kinds of problems.

The objects comprising a new type of problem (a target problem) are often dissimilar to their counterparts in a learning example of the same type (a source problem). Consequently, most instances of mapping a source problem to a target problem are new instances of mapping dissimilar objects. The declarative knowledge from yesterday's math assignment has no bearing on today's science assignment, but the procedure for mapping a source problem to a target problem is governed by the same basic set of analogical productions in either case. The compilation process that determines the productions of this general set is the same as the compilation process that determines the set of productions for a more specific skill (Anderson, 1987). In other words, there is no obvious reason why weak-method mapping should not be developed in tandem with more specific problem solving skills if a process of analogy is used to acquire many different problem solving skills. In fact, Singley and Anderson (1989) raise the possibility explicitly by suggesting that this process could be a mechanism of relatively general transfer, but dismiss the possibility only after concluding that analogy is an overlearned skill by the time adulthood is reached:

> One very important weak method in our theory is the set of productions that implement the process of analogical interpretation…This analogy process derives prescriptions for action from declarative knowledge and thus provides a bridge from declarative to procedural knowledge…The one problem with reviving [general transfer] under the banner of weak methods is that, by the time problem solvers reach adulthood, the various weak methods are already well-practiced and are a well-established part of the standard repertoire of problem-solving methods (Newell and Simon, 1972). Since the weak methods are so overlearned, they cannot serve as the basis for transfer between tasks (pp. 229–230).

The only major distinction between Singley and Anderson's (1989) conclusion and what I have proposed is my assumption that this weak method is somewhat less developed than they suggest in most people, and that it was perhaps developed less still in the past. Given the emphasis that the ACT-R theory places on the role of acquired knowledge, it is doubtful these authors would dispute that differences in performance between two populations could, at least in principle, be explained by differences in the productions that instantiate analogy if members of one population have far more experience with analogical problem solving than members of the other. A secondary distinction, if it is really a distinction at all, is my assumption that higher-level organizational knowledge (Salvucci & Anderson, 2001) can be acquired along with analogy itself if it is applicable to many types of problems. There is no reason why this should not be the case. The production compilation process is governed from top-down by demands of common tasks, and is therefore effectively blind to bottom-up theoretical distinctions such as the difference between mapping itself and how it is applied to various tasks.

If what I am suggesting is true, transferable procedural knowledge for analogy can be developed by learning specific problem solving skills that are not themselves primarily analogical. In other words, *an immediate ecological cause of the Flynn effect could be learning to solve diverse, unfamiliar problems of the sort that require an initial process of working through an example.* Working through a difficult math problem from a textbook by mapping it onto the examples at the beginning of a section is a relevant exercise that should be familiar to most readers. However, the value of this exercise for shaping procedural knowledge depends on how familiar the content is to the problem solver. Up to a point, the potential for developing weak-method mapping should be inversely related to the similarity of objects between new target

and source problems. Little benefit is conferred when problem solvers represent a new target problem's objects as similar to its source problem's objects; more of a benefit is possible when participants represent a new target problem's objects as dissimilar to its source problem's objects. Of course, too much dissimilarity will simply result in failure to map the problems and little or no acquisition of weak-method mapping.

One could argue that the example-based problem solving that I describe should not require actively testing relations because the relations between objects in most real-world problem domains are implied by the objects themselves. For example, math formulas make use of objects with familiar roles (e.g., divisor, addend, etc.), and are useful precisely because these objects occupy their ordinary roles within the formula. The problem with this argument is that it neglects the complexity of the higher-level roles that are served by combinations of familiar roles. Most students learning the Pythagorean Theorem, for example, know what all the objects "do," but even so, are unlikely to understand why the formula works unless they have been shown a proof. Nevertheless, most students learn how to apply the formula, which implies that they must rely on alternative procedures for mapping their classroom or homework problems onto examples. Those who attempt to represent relative magnitudes of distances as roles would succeed at solving the problem because c^2 must correspond to the long face of the triangle. Generating and applying an incidental relation such as *magnitude* requires no knowledge of the more abstract relations that cause the theorem to actually work. A more general relation such as *number* (e.g., number of coefficients per term in a mathematical function) would apply to many kinds of problems.

Although I have focused mostly on procedural knowledge, the previous paragraph acknowledges that development of weak-method mapping may be accompanied by either development of new roles and relational chunks in declarative knowledge, or at least greater access to those that are already available. In a recent paper, Doumas et al. (2008) demonstrated how a cognitive architecture can create new relations by consolidating distinct instances into a single relation that did not previously exist. Their model, Discovery of Relations by Analogy (DORA), is much like its predecessor LISA (Hummel & Holyoak, 1997), but begins learning with more primitive *invariant* concepts. In the terminology I have used, DORA is essentially an attempt to resolve the apparent paradox that indeterminacy presents to initial acquisition of relations: how can an architecture learn a relation from observing instances of roles when what is

deemed a role is determined by the relation (e.g., how can one generalize an instance of *larger than* without first possessing a concept of *size*)?

> Armed with a basic vocabulary of perceptual and relational invariants (which may be either present at birth, the result of specific computing modules, or some combination of both), DORA discovers relations through general learning processes and develops as a result of experience. Its development reflects a cascading process in which, through learning, initially holistic features become represented as explicit predicates, which then become more refined and get combined into relations, which themselves become more refined. The resulting qualitative changes in the model's representations—and thus in its ability to reason—reflect the operation of basic learning processes generating more sophisticated representations by building on previously acquired representations (Doumas et al., 2008, p. 30).

Doumas et al.'s (2008) work is suggestive that learning to solve many diverse problems could raise subsequent probabilities of retrieving widely applicable roles and relations. Given the developmental interplay between procedural and declarative knowledge, is probable that the procedural weak-method I have proposed has a declarative component as well.

The potential role of example-based problem solving in explaining the Flynn effect deserves more consideration than I can devote here. Although I have used math examples and review evidence below that implicates changes in math curricula as a source of example-based problem solving (Baker et al., 2010), this should not be interpreted as an implicit claim about the content that relevant problems must possess. My only claim is that content must be diverse enough for new instances of example-based problem solving to be novel. Moreover, I do not claim that example-based problem solving is the exclusive cause of improved mapping. For example, it should also be possible to improve mapping by solving problems where analogy or induction is itself the primary goal as opposed to a mere means by which instructions and practice problems are transformed into procedures. As noted, multiple studies show that inductive reasoning training improves scores on tests (Klauer & Phye, 2008). After applying the theory to other tests, I review evidence below suggesting that tasks demanding analogical reasoning did begin to appear in elementary math texts (as exercises to promote use of multiple commensurable

strategies when solving a single type of problem) in the second half of the 20th century (Baker et al., 2010).

Generalization to Other Tests

The gold standard for any theory of rising scores is accounting for gains on the Ravens and similar tests with the largest Flynn effects. In fact, emphasis on the abstract Ravens has led me to understate the impact weak-method mapping could have on other tests that are seemingly less abstract. Consider the types of items on the WISC Similarities, a test for children with a very large Flynn effect (Flynn, 1999; Flynn & Weiss, 2007). The following example is adapted from Flynn and Weiss (2007), who note that Similarities appears to be distinct from tests like the Ravens due to its verbal content and its superficial appearance as a test of acquired knowledge (it is considered a test of verbal ability). Similarities requires children to compare two analogs such as *dusk* and *dawn.* Answers based on surface similarities such as *time of day* or *intermediate brightness* (however they may actually be verbalized by a child) would receive lower scores than answers based on deeper similarities such as *separates night and day* (Flynn & Weiss, 2007). Assuming that children are familiar with dusk and dawn, concurrent presentation of these two concepts would elicit others that are common to both such as the examples above. *Time of day* and *intermediate brightness* are common objects and roles that may be retrieved spontaneously and offered indiscriminately by a child who does not test for deeper relations. However, weak-method mapping makes it possible to generate and evaluate further possibilities. Assuming a skilled problem solver retrieves both *time of day* and *intermediate brightness*, she is at least capable of representing them as objects in need of roles. That she will is not ensured by the productions (this depends on other organizational knowledge, including, but not limited to, task instructions) but if she does, she is capable of inducing the relation, *separates night and day*. The major difference between her and an unskilled problem solver is that she is flexible enough to treat a prospective role as an object in need of a role if the situation demands. This does not imply that she will always do so when appropriate, nor that she would not benefit from additional organizational knowledge (e.g., a simple and general heuristic of attempting to account for the most objects with the fewest relations). A greater facility for treating roles as objects can help to explain why today's average child scores at the 94th percentile of her grandparents' generation on Similarities (Flynn & Weiss, 2007). There is no reason why a weak-

method for mapping dissimilar objects should disappear in the presence of content. In this case it could lead to higher scores on a test designed to assess verbal ability.

It is tempting to assume that weak-method mapping would have no effect on tests that do not test analogical reasoning explicitly. However, several decades of research by K. J. Klauer and his colleagues (e.g., Klauer, 1996; Klauer, Willmes, & Phye, 2002) suggests that training children to search for similarities (the implication being that children must learn that similarities are not always readily apparent) imparts a surprisingly generalizable skill. Klauer and Phye's (2008) meta-analysis of 74 experiments, consisting of nearly 4,000 children, revealed that a form of inductive reasoning training using "meaningful material and…problems that children may encounter in their daily lives" (Klauer & Phye, 2008, p. 93) has non-trivial effects that generalize beyond inductive reasoning tests. Remarkably, the effect on learning of an academic topic (assessed with pre- and post-tests) was found to be larger than the effect on fluid ability tests (weighted mean effect size of d = .69 and d = .52 respectively; the difference between theses effects is statistically significant), which implies that weak-method mapping could raise scores on tests that are not purely analogical or inductive.

The question is how knowledge that is specific to analogy could affect responses to non-analogical items. An intriguing possibility is that it enhances mapping of objects between *whole items.* Regardless of the targeted ability, most intelligence tests feature multiple items of varying difficulty that have a very similar basic structure; that is, items are analogs to one another. In fact, mapping one item to another on a test without analogies is more like example-based problem solving than is solving items that are analogies themselves. Ravens items are all analogically similar to one another, as are items on spatial ability tests, verbal ability tests, and many others. Weak-method mapping could make it possible to generalize a rule from one item, such as an abstract goal state or an entire solution strategy, to other items on the same test even though each item has distinct content. In discussing the effect of instructions, I have already shown how this similarity could improve scores of individuals whose initial problem representations are flexible enough to apply to every item.

If the ability to map objects between items has contributed to higher scores, gains should be largest for tests composed of items with a structure that is both initially unfamiliar and relatively uniform from item to item. Weak-method mapping would confer little or no advantage on tests with structures that are highly familiar to test-takers, or tests composed of items that are not

analogically similar to one another. The popular Wechsler and Stanford-Binet tests both consist of many subtests requiring problem solving procedures that would seldom be encountered outside the context of intelligence testing, and that remain relatively consistent throughout any individual test. Consistent with predictions, the tests show moderate improvement overall, with the lowest gains on subtests consisting of items that are amenable to the same procedures that apply to schoolwork or scholastic achievement tests, such as Arithmetic, Information (a test of general knowledge), and Vocabulary (Flynn, 1999; Flynn & Weiss, 2007). There is little to be gained from mapping objects between items on these subtests because their structures are already familiar to every test-taker. Even if they were unfamiliar, the items demand declarative knowledge that must be acquired prior to the test. In contrast, subtests bearing little resemblance to traditional schoolwork such as Similarities, Picture arrangement, Block assembly, and Coding show considerably larger gains (Flynn, 1999; Flynn & Weiss, 2007). These subtests have problem structures that are relatively uniform throughout and unfamiliar to most test-takers.

In general, the theory predicts that gains in raw scores should be highest on tests where analogical reasoning is most crucial, regardless of whether the tests were designed to assess this ability or not. How participants obtain solutions to items is a question of the actual goals and sub-goals they must accomplish to solve an item that can only be answered by conducting a task analysis (e.g., Ericsson & Simon, 1993).

Isolating Genuine Ecological Causes: The First Hundred Years of "New Math"

At the outset I suggested that identifying a cognitive mechanism could help to narrow the field of potential ecological causes. When discussing causes of a trend (e.g., increasing rate of obesity) it is important to distinguish between ultimate causes on the one hand (e.g., television), and proximal causes on the other (e.g., recreational eating and inactivity). Because this is a cognitive theory, I restrict consideration to proximal causes. I have already dismissed genetic changes and improved nutrition as major contributors, but it is still necessary to distinguish between multiple knowledge-based explanations. Results do not permit dismissing any knowledge-based explanations outright, but they do appear to implicate one prospective cause in particular.

It is seldom mentioned that mean IQ gains have not tended to accrue within individuals over time, much less at a pace commensurate with between-cohort gains. IQ scores rose at a mean rate of 3 to 6 points per decade (Flynn, 2007) for every new cohort or cross-section affected, but the

average individual has not gained 3 to 6 IQ points per decade of life. Why should this be? That IQ is stable over the lifespan is just a restatement of the question, which is essentially, what about infancy or early childhood plays such a consequential role in determining test scores over the course of one's life. The most obvious answer is education, which impacts children at an early age, and more relevantly, is a major source of the example-based problem solving exercises that should stimulate development of weak-method mapping.

Several investigators have argued for education as a proximal cause (Ceci, 1991; Flynn, 2007; Flynn & Weiss, 2007, Blair et al., 2005), but surprisingly few studies are relevant to understanding the role of education in the development of abstract reasoning. As Ceci (1991, p. 715) lamented, "although one might imagine that the one area in which there would be great activity among school researchers is that which deals with the relationship between schooling and conceptual development, this has not been the case." Ceci's (1991, p. 715) exhaustive review led him to conclude that schooling raises IQ scores above what they would be in the absence of formal education, and suggested that development of conceptual skills may be one of the reasons why:

> Compared with their nonschooled peers, schooled children are more likely to (a) sort stimuli by form and class rather than by color, (b) group items that belong to the same taxonomic class rather than to the same thematic class, (c) demonstrate greater flexibility in shifting between domains during problem solving, and (d) spontaneously engage in more verbal descriptions of their classifications (Ceci & Liker, 1986a; Evans & Segal, 1969; Gay & Cole, 1967; Greenfield, Reich, & Olver, 1966; Hall, 1972; Irwin & McLaughlin, 1970; Schmidt & Nzimande, 1970; Sharp, Cole, & Lave, 1979; Stevenson et al., 1978).

Although Ceci (1991) mentions other skills, such as perception and memory, variation within cohorts is really only relevant to the Flynn effect if it can be linked to hypothetical sources of between-cohort variation. The skills cited above are expected consequences of observed changes in educational practices (Flynn, 2007; Blair et al., 2005; Flynn & Weiss, 2007). Following an analysis of textbooks, Blair et al. (2005) concluded that

> at the turn of the 20th century, much of the mathematics instruction for children in the upper elementary grades was rigid, formalistic, and emphasized drill and rote memorization (p. 99) [but today] young children regularly engage in visual-spatial

> problem solving …that their grandparents' generation would not have been exposed to until the seventh or eighth grade and that their great-grandparents' generation may not have been introduced to at all (pp. 101–102).

They buttressed their claim with examples from textbooks published over the century that show abandonment of formal rigid exercises in favor of problems emphasizing abstract reasoning, often in the absence of numbers or other familiar symbols. A subsequent analysis of textbooks by Baker et al. (2010) merits close consideration for being the most thorough and systematic analysis of its kind.

Concerned that lack of empirical evidence on curricular content and changes in content over time has forced education policy makers to rely on highly speculative assumptions, Baker et al.'s (2010) goal was to "provide thorough empirical evidence of content in elementary school mathematics textbooks in the United States over the course of the 20th century" (p. 384). The authors collected 141 widely-used kindergarten through 6^{th}-grade math textbooks published between 1904 and 2000, with a combined total of over 28,000 pages. They classified the content of individual pages into six categories of basic arithmetic, advanced arithmetic, geometry and measurement, reasoning based in formal mathematics, reasoning not based in formal mathematics, and miscellaneous material.

According to their analysis, content for most of the first half of the century, although challenging, was relatively narrow and not very abstract. For example, "earlier textbooks often provided a page of many facts, such as 12 + 5 = 17, 12 + 6 = 18, and 12 + 7 = 19, and then asked students to review the problems repeatedly until they could recite them from memory under a specific time constraint" (p. 408). Such an exercise would have virtually no impact on weak-method mapping, at least for students who are familiar with the concepts denoted by these symbols. Reasoning (with or without formal mathematics), which the authors define as "pattern solving, informal algebra, informal geometry, probability, and grouping/categorization" (p. 413) suddenly began to increase dramatically after mid-century:

> Until the mid-1960s this kind of content was rare, covering barely 3% (2 pages) of first- through third-grade textbooks and at most 5% (5 or 6 pages) of fourth- through sixth-grade textbooks. Yet by the end of century, first- through third-grade students used textbooks that had on average 12% (2 to 39 pages…)

> reasoning content and fourth- through sixth-grade students learned from textbooks that had 17% (5 to 60 pages…) reasoning content (pp. 399–400).

It is noteworthy that the number of pages in textbooks more than tripled during the century, increasing from around 100 pages in 1904 to well over 300 in by the end of the century across grades, meaning that percentage of pages understates the magnitude of trends. By the absolute standard of pages, children at any given time were exposed to greater quantities of reasoning material that encompassed a wider breadth of content than their predecessors had at the same grade level even relatively early in the century. All told, an average child in the year 2000 used a textbook with roughly *40 to 60 times* more pages of reasoning content than a child in 1904.

There were several other important trends that emerged only after mid-century, including an increase in the number of topic areas covered in textbooks (resulting in an increase of 50 percent for the whole century), lower age of exposure to abstract conceptual material, and a related increase in the sudden presentation of new material without precedent. For example, informal geometry (classified as reasoning because the problems generally do not have algorithmic solutions) was virtually absent from the books for any grade level published prior to the 1970s, but then began to appear randomly as early as kindergarten, accounting for nine percent of the text in one kindergarten book published in 1991. This is even more remarkable when one considers that "in the first decade of the 1900s, mathematics was rarely taught before the second grade" (p. 400). Interestingly, Baker et al. (2010) observed greater emphasis on learning multiple problem solving strategies for the same problems beginning in the 1960s, and later, emphasis on understanding the conceptual relationships that make multiple strategies commensurable:

> Comparatively, textbooks from the mid-1960s onward presented more problem solving strategies requiring the use and mastery of field properties for real numbers (often in the context of whole numbers for the youngest students) and/or the application of multiple previously learned strategies to achieve a new and deeper understanding. Although these later textbooks continued to include the types of strategies found in textbooks from the first three historical periods, additional new problem solving strategies required students to use field properties (e.g., commutative property of addition for real numbers) to expand their existing knowledge or decompose more difficult problems into familiar ones (p. 409).

Although Baker et al.'s (2010) findings are too complex to be translated into any simple take-home message, they are strikingly compatible with the current proposal. The trends these authors report are characterized by a profusion of learning exercises that call for example-based problem solving. With that said, it is potentially misleading to generalize trends in American education to the rest of the developed world where American educational practices could be regarded, quite reasonably, as a model of what does not work. A series of cross-national analyses similar to Baker et al.'s (2010) would contribute enormously to current understanding of how education shapes the cognition of abstract analogy. Given the information available, it is still too early to rush to conclusions, but I suggest that the evidence for education as a proximal cause is strong enough to merit a thorough analysis of curricular changes before committing to more exotic hypotheses. One could argue that what may be at stake outweighs the time and cost of pursuing a false lead.

Evaluating the Theory

Although I have already alluded to several ways of testing the theory, in this section I outline more specific predictions. These predictions are decisive, but the cost of this decisiveness is the need to consider some meta-theoretical issues that are unavoidable when positing hidden mechanisms or processes.

Prospective Tests

The broadest and most easily tested predictions are simple extrapolations of the core assumption that analogical mapping is an acquired skill and that mapping dissimilar objects is more difficult than mapping similar objects. To the extent that test scores rise over time within a population, individuals from later cohorts will tend to be more successful at mapping objects. However, it is important to consider that the nature of the difference in mapping between two cohorts is also expected to change as a function of time period and magnitude of difference in scores. Studies 1 and 2 tested simple predictions because the overall differences in scores between cohort 1 and 2 were relatively small in both cases. Different predictions would have been appropriate if, for example, Study 1 had compared contemporary younger adults to their counterparts from around 1920 because the theory assumes few young adults in 1920 would be especially good at mapping even relatively similar objects. Such a study would be uninformative

because verifying the prediction of little or no correlation between change in pass rate and dissimilarity would not rule out alternative explanations.

This is why testing the most decisive predictions requires designing sets of new items that differ as systematically and unambiguously as possible with respect to similarity of objects (e.g., Primi, 2002). Theoretical guesswork can be minimized by constructing items out of features that virtually any person in any population would be expected to represent as objects, at least to the extent that this is possible. New data can be obtained from developing countries where Flynn effects are most likely to be occurring at the present time.

As stated, matrix reasoning tests are not the only tests to show Flynn effects, nor are they the only tests to demand analogical reasoning. The theory can and should be applied to other tests. Unfortunately, there can be no standard-operating-procedure for classifying test or item properties other than identifying corresponding objects and classifying them according to the abstractness of their roles. The decisiveness of findings will be compromised when the analogical components of tests do not lend themselves to clean classifications; however, I suggest below that the theoretical rationale behind the system used in this paper (illustrated in Table 2) is sound justification for applying the same principle to other tests.

In recent years, innovative psychometricians have increasingly recognized the need for new item response models capable of accommodating cognitively plausible theories of item responses (De Boeck & Wilson, 1994; Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). The assumptions of Studies 1 and 2 that number of rules and dissimilarity of objects are factors in some kind of constant-conjunction system of causality make these studies interpretable in light of previous work, but oversimplify the item-response process. An item with one rule containing dissimilar objects should be nearly as difficult as an item with four rules containing dissimilar objects for someone who cannot map dissimilar objects. Study 2 makes use of binary outcomes in a relatively simple model only because the verbal report methodology makes it possible to observe processes almost directly. Application of this same model or a similar Rasch model to item responses would entail making tacit quantitative assumptions about *whole sequences of processes* that are cognitively implausible and incompatible with the theory. Developing models that formalize process theories appropriately is a challenge, but one that cannot be avoided if the goal is to understand the psychological causes of performance and performance variation.

The proposal that example-based problem solving is a transfer mechanism for analogical mapping affords some additional predictions. In fact, testing these predictions could shed additional light on both the Flynn effect specifically and transfer in general. The basic idea is that learning to solve abstract problems can be an analogical training task in its own right when learning by example is necessary, and when the types of problems are diverse and unfamiliar to participants. Conversely, learning to solve problems should not be sufficient all by itself to improve analogical reasoning, which implies that learning to solve problems that do not require working through an example, or learning to solve many versions of a single type of problem that does should not be very beneficial. Training materials should be selected in accordance with the competencies of the target population. For example, individuals with relatively little exposure to abstract problem solving tasks should have the most to gain from learning a variety of diverse problem solving skills. However, improvement requires successfully mapping objects from problems of a new type to objects of examples, which requires at least some measure of relevant prior knowledge. Ideally, the mapping goal should be difficult but achievable.

Experimental training studies can provide strong evidence either for or against example-based problem solving as a cause. However, the standard short-term transfer design (a one-session experiment) is not suitable for rejection (see Anderson, Reder, & Simon, 1996). The procedural knowledge in question develops in response to solving a variety of problems of various kinds over a period of years in the real world, and while it should be possible to accelerate its acquisition in a dedicated training context, a null transfer effect observed after a one or two hours of training is not informative. Findings are only valuable to the extent that the method used to obtain them does not defeat the purpose of the investigation.

Additional Considerations

The prospective studies described above are intended to test the theory as presented. However, additional questions remain about the so-called auxiliary assumptions that were made in order to test the theory. These questions are primarily representational: how do participants represent objects, how do they represent the task (or task*s* for those who do not have a generalizable representation), and how uniform are procedural representations of two participants with the same level of skill?

Two otherwise identical architectures can conceivably represent objects differently from one another, in which case the same productions can be expected to elicit different item responses.

The choice to equate perceptual similarity with literal identicalness of objects (with a few exceptions) was mostly pragmatic as this definition is relatively simple and is expected to correspond reasonably well to the actual criteria by which participants decompose the figures. However, intuition suggests this definition is problematic as it applies to certain items. I already mentioned Figure 5, which is isomorphic to item 27 on the Ravens. Like Figure 5, item 27 consists of three versions of three basic shapes (triangle, circle, and rectangle). Not surprisingly, many low performers and cohort 1 participants in Study 2 were able to infer that the correct answer must contain a circular shape (unlike Figure 5, which calls for a rectangle). Although results like this contradict the formalization of the theory for empirical testing, it is less clear that they signal any major theoretical flaw beyond their revelation that improved criteria are needed for determining initial representations of objects. Ideally, these criteria should be informed by theories of perception and categorization, and how these processes are influenced by culture.

Questions of object representation extend to the dissimilarity variable used in Studies 1 and 2. The variable was created based on properties of the Ravens and cannot be translated wholesale to other tests. However, the principle of abstraction underlying the variable—abstract roles with dissimilar objects subsume concrete roles with similar objects—is sufficiently general to be applied to other tests by decomposing analogs (be they parts of items, whole items or something else) into constituent roles and relations. Note that the highest level will, almost by necessity, correspond to properties of other tests for the same reason that it is the highest level with respect to the Ravens. The most abstract roles and relations are not confined to any single type of problem because they do not refer to concrete features. As long as the application of the principle to other tests is consistent, transparent, and coherent to others, the present work supplies theoretical justification for comparing findings from distinct tests.

The productions are presented at a grain-size that combines different hypothetical production lists from different individuals into a single category of lists that serve the same function. This lack of specificity should be evaluated in light of the realization that the mechanisms characteristic of a single individual cannot fully capture the nature and magnitude of variation in one or more populations.[11] I contend that using a larger grain-size to explain a group difference

[11] Much could be learned from a research program devoted to simulating populations of persons with populations of cognitive architectures whereby each architecture is formalized for multiple tasks, ideally with the same free

is justified so long as the theory remains specific enough to make behavioral predictions that are distinguishable from those made by alternative theories. In the present case, the set of possible productions is sufficiently constrained to predict responses, verbalizations, eye movements, and other observable behaviors as a function of cohort and item or task properties. Singley and Anderson (1989) address a similar concern in the final chapter of their book with an argument that applies here:

> The reader should not leave this chapter with the conclusion that representational issues [choice of productions] pose a particular problem for the study of transfer. They are problems for the study of all cognitive phenomena, and for the same reason: there is the danger that the theorist can salvage a mistaken theory by suitable choice of representation. If anything, transfer is less imperiled by representational indeterminism than are most phenomena because so many constraints can be applied to the representation before making behavioral predictions (p. 274).

What I have proposed is a departure from convention, but contains few if any revolutionary claims about skill acquisition, transfer, or analogical reasoning. It is not the theory, but rather the application of a skill acquisition framework to test scores that is unconventional. By presenting two findings that permit no radically different explanation, I have made a plausible case for changes in analogical mapping as a cognitive cause of the Flynn effect. The cognitive literature does not appear to offer any alternative conceptualization of learning that is flexible enough to account for higher scores on abstract tests without appealing to learning exercises that more closely resemble the tests themselves. Of course, the apparent absence of a vastly different alternative does not rule out the possibility that one exists, and even if no such alternative is forthcoming, the additional possibility remains that there are multiple cognitive causes of rising scores, no single one of which accounts for a disproportionate share of the effect. If so, the conceptualization of transfer as common productions remains an ideal framework for considering additional mechanisms. I conclude by arguing that a bottom-up approach to theorizing (Borsboom et al., 2004; Cervone, 1997), including but not limited to concepts like productions, is ultimately essential to comparing populations.

---

parameters held constant. This would help to address (and better illustrate) the problem I consider in the final section.

A Bottom-up Approach to Psychological Generalization

I stated earlier that the message of this paper cannot be understood without suspending the interpretation of between-subjects variation as a causal entity in its own right, but did not fully articulate why doing so is essential to making sense of cross-cultural anomalies such as the Flynn effect.

Observed constructs and latent variables (constructs) are often portrayed as intrinsic, causal psychological properties of human beings. Yet, much like ordinary operational variables, constructs are relational to population (Borsboom et al., 2004; Lord & Novick, 1968) and have no comprehensible interpretation at the level of the individual (Lamiell, 2011). This is because they are *differences between persons*, which are distinct from properties of persons (Lamiell, 2007) in that they distinguish one person from another rather than characterize individual persons in isolation. It makes perfect sense to speak of the height of an individual in isolation, but it is nonsensical to talk about the intelligence or extroversion of an individual without implicit reference to some other person or group. The problem is that any category that varies *by definition* can have only variation itself as a referent, which is why any generalizable psychological explanation must refer to something that is at least conceivable in a population of one person (see Van der Maas et al., 2011).

Now consider that the psychological processes that determine item responses (Borsboom et al., 2004; Snow and Lohman, 1989) are logically independent of relative performance within populations. It is possible to describe one person's solution to a highly discriminating test item as an outcome of some series of processes without invoking any concepts *defined by* success at solving items of that kind relative to other persons.[12] Thus, a theory of the psychological processes that determine observed performance can be a reference point for comparing constructs or observations in two distinct populations (see Borsboom et al., 2003, 2004). Such a theory must be generated from the bottom up (Borsboom et al., 2004; Cervone, 1997), and refer to properties of persons rather than aggregates (Van der Maas et al., 2011) to sustain the logical independence of observation and cause.

[12] The argument, to be clear, is not that processes are intrinsic properties, but rather that they are logically independent of lexical categories that denote differences between persons. My discussion of the grain-size of productions can be read as acknowledgement that what is considered a process depends on the investigator's decomposition of a task.

The preceding paragraphs should be read not as criticism of prevailing methods (see Olsen & Morgan, 2005) but rather as affirmation of the distinction between psychological causes and methodological constructions such as covariation (Borsboom et al., 2004). The inherent limitations of construct validation theory (Cronbach & Meehl, 1955) are implicit in Bakan's (1955) contemporaneous observation, which remains as relevant today as it was sixty years ago:

> The failure to distinguish between general-type and aggregate-type propositions is at the root of a considerable amount of confusion which currently prevails in psychology…A general-type proposition asserts something which is presumably true of each and every member of a designable class. An aggregate-type proposition asserts something which is presumably true of the class considered as an aggregate (p. 211)…The distinction between the two types of propositions does not preclude the possibility of using one type of proposition as a basis for inference with respect to the other type. However, this is quite different from the syncretic substitution of inappropriate research methods (p. 212).

Conclusion

The present work is an effort to explain rising scores on culture-free intelligence tests as a knowledge-based phenomenon. The work applies Singley and Anderson's (1989) theory of transfer as common productions to Flynn's proposal that rising scores were caused by improved abstract reasoning cultivated by emphasis on scientific thinking. A review of the literature suggests that the ability to map dissimilar objects is a source of variation in scores on culture free tests, and a study of archival data shows that contemporary young adults are better at mapping dissimilar objects than their counterparts were fifty years ago. A process tracing study provides further evidence that differences in mapping of dissimilar objects are a cause of differences in scores between younger adults of today and younger adults of fifty years ago in scores achieved on the test with the largest Flynn effect. It is notable that neither finding appears to be compatible with any other known explanation.

To the extent that the theory is accurate, rising test scores, at least on culture-free tests, are caused by development of a weak method for analogical mapping of objects, realized in the form of productions, and perhaps, greater access to chunks representing common roles and relations. Previous work suggests that a weak method could develop in response to solving diverse, unfamiliar problems that require an initial process of working through examples. The theory

makes specific, testable predictions by identifying the features of items that render them sensitive to improved abstract analogical reasoning. Although it is too early to conclude that an analogy-specific mechanism can explain the entirety of the Flynn effect, results imply that such a mechanism is at least a plausible starting point. In either case, the conceptualization of transfer as procedural knowledge is an appropriate framework for generating, advancing, and testing supplementary or alternative mechanisms.

In addition to its relevance to the Flynn effect, the work furthers understanding of the cognition of matrix reasoning. Results of both studies suggest that cohort or time period contribute substantially to cross-sectional findings that are often attributed to age-related mechanisms, and Study 2 in particular highlights the possibility that generalization of instructions across items is a skill-dependent phenomenon.

The most visible contribution of the work is its relevance to understanding why the very tests designed to minimize the influence of culture were so susceptible to cultural changes. Removal of familiar content could minimize the role of knowledge as it is traditionally conceived, but could not eradicate culture as a determinant of scores because the problem structure remained as the context for a subtler form of knowledge. This procedural knowledge is also bound to culture, but is by its very nature scarcely conceivable within a theoretical framework that can sustain culture-free intelligence as a unitary concept. In 1930, some eight years prior to publication of the preeminent culture-free test, the gestalter, Wolfgang Köhler, expressed skepticism that observable differences between persons correspond to categories that are psychologically significant in an absolute sense:

> By standardizing against scholastic or occupational performance, we call *intelligence* in our children something which merely corresponds to those particular requirements which the present-day school, the present-day city in Europe or America, the present-day middle class consider important. But how do we know that something so arbitrarily taken out of history, geography, and all cultural possibilities can serve as a suitable measure of a basic [i.e., general-type as opposed to aggregate-type] psychological attribute? And yet we can scarcely proceed otherwise if the test is not based on an analysis of the decisive processes involved (as translated in Henle, 1971, p. 187; original emphasis).

If the Flynn effect is a testament to the capacity of humans to adapt to their environments, then it is also a statement about the vastness and *irregularity* of human diversity. The need to accommodate this irregularity will become increasingly apparent as cross-cultural, cross-geographical findings accumulate in the coming years (see Henrich, Heine, & Norenzayan, 2010). Establishing a psychology that can cope with diversity and change will require looking beneath the surface features of human variation for principles that transcend both culture and time.

APPENDIX A

EXAMPLES OF VERBALIZED CORRECT OBJECTS

Level 1

Item 1: "I think it would be three lines through."

Item 5: "So I'm looking for almost an L."

Item 13: "Yeah find something with three lines."

Item 17: "It's actually going to be a full triangle on the end."

Item 24: "And then vertical lines down top right."

Item 25: "It's gonna be all stripes and…"

Item 32: "Line line on the bottom."

Level 2

Item 8: "Vertical, the vertical part is colored in.

Item 12: "So just a dot and the only one that could be is six"

Item 13: "And the upright rectangle should appear in the bottom right…"

Item 15: "You're gonna have to make that um a wavy one."

Item 16: "Or would it be just a circle?"

Item 17: "so it's gotta be straight so its number six."

Item 18: "Have to switch the angle."

Item 21: "It goes with the bow tie."

Item 31: "A diagonal then all the way to the right."

Item 34: "Um it's the bumpy pattern and the squiggles…"

Level 3

Item 22: "…which is a square with a circle not with the dot."

Item 23: "That would have four dots and a circle I believe."

Item 25: "The next one will be diagonal with a half moon to the bottom right."

Item 26: "The thingy is going to point to the right."

Item 27: "It might be a paper fold."

Item 31: "So in the third row there should be a blank on the left hand side."

Item 32: "Nothing at the top right hand corner."

Item 34: “And I think I’ll take a look at the c and it goes like that.”

APPENDIX B

HUMAN SUBJECTS APPROVAL LETTER

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 Â· FAX (850) 644-4392

APPROVAL MEMORANDUM (for change in research protocol)

Date: 9/24/2008

To: Mark Fox [redacted]

Address: 1107 W. Call St. Tallahassee FL 32306 or Department of Psychology
Dept.: PSYCHOLOGY DEPARTMENT

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research (Approval for Change in Protocol)
Project entitled: Concurrent Verbalization and Cognitive Performance

The form that you submitted to this office in regard to the requested change/amendment to your research protocol for the above-referenced project has been reviewed and approved.

Please be reminded that if the project has not been completed by 6/18/2009, you must request renewed approval for continuation of the project.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Neil Charness, Advisor [charness@psy.fsu.edu]
HSC No. 2008.1659

APPENDIX C

APPROVED HUMAN SUBJECTS CONSENT FORM

INFORMED CONSENT FORM

Date:_____

I freely and voluntarily consent to be a participant in the research project entitled "Concurrent Verbalization and Cognitive performance". Mark Fox, Ainsley Mitchum, and Dr. Neil Charness will be the principal investigators.

I understand that I will be given tests measuring different cognitive abilities. Abilities that will be tested are inductive reasoning and task-switching skills. I will also answer questions soliciting my personal views about the nature of intelligence and how I prepare for challenging tasks. In addition, I understand that I may be observed during a typical session and that this session could be audio taped to capture talk/think-aloud information (e.g. speaking my thoughts aloud while I perform) for later protocol analysis. After completion of a task, a retrospective recording of my thoughts during performance may also be used for later analysis. I may be outfitted with a headset that tracks my eye-movements so that that specific information about my visual attention can be monitored. I understand that this experiment will last approximately one to one and a half hours.

I understand that the experimental tasks do not present more risks than people encounter in everyday life, doing academic work or solving puzzles. I may become tired at some point, and I may ask the experimenter to take a short break between tasks. The benefit is that I will either receive course credit toward my research participation if I am a student, or $10 per hour if I am an older adult. I will also be given a lesson about the experiment at the end so that I may learn about cognitive processes.

I understand that the records of this research which refer to my data will be given a code so that no one except the investigators and their designated assistants will have access to the data, and that no identifiable data, including handwritten information that I have supplied, will be used for publication. In addition, the records of this research, which refer to my performance, will be kept confidential to the extent allowed by law. I understand that any audio tapes used in this project will be retained at the FSU Department of Psychology, and that the tapes will be erased or destroyed within ten years (September or 2018). I understand that I will be paid 10 dollars per hour for participation in this project if I am not a student at Florida State University and one credit per hour if I am a student at Florida State University.

I understand that I may stop the experiment at any time without penalty. I understand that I may contact [redacted] to questions about this research. I have read and I understand the foregoing.

If I have questions about my rights as a subject/participant in this research, or if I feel that I have been placed at risk, I can contact the Chair of the Human Subjects Committee, Institutional Review Board, through the Office of the Vice President for Research at (850) 644-8633.

Signature of Research Participant_____

Printed Name_____

FSU Human Subjects Committee Approved on 6/19/2008. Void after 6/18/2009. HSC#: 2008.1312

REFERENCES

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem situations. *Psychological Review, 94,* 192–210. doi: 10.1037/0033-295X.94.2.1

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1322–1340. doi: 10.1037/0278-7393.20.6.1322

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limits on retrieval. *Cognitive Psychology, 30,* 221–256. doi: 10.1006/cogp.1996.0007

Anderson, J. R., Reder, L.M., & Simon, H.A. (1996). Situated learning and education. *Educational Researcher, 25,* 5–11.

Babcock, R. L. (1994). Analysis of adult differences on the Raven's Advanced Progressive Matrices Test. *Psychology and Aging, 9,* 303–314. doi: 10.1037/0882-7974.9.2.303

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's Advanced Progressive Matrices. *Intelligence,* 30, 485–503. doi:10.1016/S0160-2896(02)00124-1

Bakan, P. (1955). The general and the aggregate: A methodological distinction. *Perceptual and Motor Skills, 5,* 211–212. doi: 10.2466/PMS.5.7.211-212

Baker, D., Knipe, H., Collins, J., Leon, J., Cummings, E., Blair, C., & Garnson, D. (2010). One hundred years of elementary school mathematics in the United States: A content analysis and cognitive assessment of textbooks from 1900 to 2000. *Journal for Research in Mathematics Education, 41,* 383–423.

Barber, N. (2005). Educational and ecological correlates of IQ: A cross-national investigation. *Intelligence, 33,* 273–284. doi: 10.1016/j.intell.2005.01.001

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 36,* 209–231.

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence, 33,* 93–106. doi: 10.1016/j.intell.2004.07.008

Block, N. (1995). How heritability misleads about race. *Cognition, 56,* 99–128. doi: 10.1016/0010-0277(95)00678-R

Boag, S. (2011). Explanation in personality psychology: "Verbal magic" and the five-factor model. *Philosophical Psychology, 24,* 223–243. doi: 10.1080/09515089.2010.548319

Bors, D. A., & Vigneau, F. (2001). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences, 13,* 291–312. doi: 10.1016/S1041-6080(03)00015-3

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219. doi: 10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Breslow, N. E., & Clayton, D. G. (1993). Appropriate inference in generalized linear models. *Journal of the American Statistical Association, 88,* 9–25. doi: 10.2307/2290687

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences, 19,* 330–338. doi: 10.1016/j.lindif.2008.10.006

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97,* 404–431. doi: 10.1037/0033-295X.97.3.404

Carriere, K. C. (2001). How good is a normal approximation for rates and proportions of low incidence events? *Communications in Statistics-Simulation and Computation, 30,* 327–337. doi: 10.1081/SAC-10000237

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27,* 703–722. doi: 10.1037/0012-1649.27.5.703

Cervone, D. (1997). Social-cognitive mechanisms and personality coherence: self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science, 8,* 43–50. doi: 10.1111/j.1467-9280.1997.tb00542.x

Charness, N., Milberg, W., & Alexander, M. P. (1988). Teaching an amnesic a complex cognitive skill. *Brain and Cognition, 8,* 253–272. doi: 10.1016/0278-2626(88)90053-X

Chi, M. T. H., & Ohlsson, S. (2005). Complex declarative learning. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 371–399). New York: Cambridge University Press.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428. doi: 10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior,* 8, 240–247. doi: 10.1016/S0022-5371(69)80069-1

Colom, R., Lluis-Font, J. M., & Andrés-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence, 33,* 83–91. doi: 10.1016/j.intell.2004.07.010

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Colored Progressive Matrices for primary school aged children in Australia. *Personality and Individual Differences, 39,* 647−660. doi: 10.1016/j.paid.2005.02.015

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281−302. doi: 10.1037/h0040957

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science, 25,* 315–353. doi: 10.1016/S0364-0213(01)00039-8

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan Children. *Psychological Science, 14,* 215–219. doi: 10.1111/1467-9280.02434

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108,* 346–369. doi: 10.1037/0033-295X.108.2.346

Dickinson, M. D., & Hiscock, M. (2010). Age-related IQ decline is reduced markedly after adjustment for the Flynn effect. *Journal of Clinical and Experimental Neuropsychology,* 32, 865–870. doi: 10.1080/13803391003596413

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115,* 1–43. doi: 10.1037/0033-295X.115.1.1

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396. doi: 10.1037/1082-989X.3.3.380

Ericsson, K. A., & Fox, M. C. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011). *Psychological Bulletin, 137,* 351–354. doi: 10.1037//a0022388

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245. doi: 10.1037/0033-295X.102.2.211

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251. doi: 10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* (Rev. ed.). Cambridge, MA: MIT Press.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51. doi: 10.1037/0033-2909.95.1.29

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191. doi: 10.1037/0033-2909.101.2.171

Flynn, J. R. (1992). Cultural distance and the limitations of IQ. In J. Lynch, C. Modgil, & S. Modgil (Eds.), *Education for cultural diversity: Convergence and divergence* (pp. 343–360). London: Falmer Press.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20. DOI: 10.1037/0003-066X.54.1.5

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect.* New York: Cambridge University Press.

Flynn, J. R., Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing, 7,* 209–224. doi: 10.1080/15305050701193587

Forbes, A. R. (1964). An item analysis of the Advanced Matrices. *British Journal of Educational Psychology, 34,* 223–236.

Fox, M. C., & Charness, N. (2010). How to gain eleven IQ points in ten minutes: Thinking aloud improves Raven's Matrices performance in older adults. *Aging, Neuropsychology, and Cognition, 17,* 191–204. doi:10.1080/13825580903042668

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137,* 316–344. doi: 10.1037//a0021663

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science: A Multidisciplinary Journal, 7,* 155–170. doi: 10.1111/j.1467-9280.1996.tb00366.x

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81–123). Washington, DC, US: American Psychological Association.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics, 56,* 915–921. doi: 10.1111/j.0006-341X.2000.00915.x

Henle, M. (Ed.). (1971). *The selected papers of Wolfgang Köhler.* Liveright: New York.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. doi: 10.1017/S0140525X0999152X

Hiscock, M. (2007). The Flynn effect and its relevance to neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 29*, 514–529. doi: 10.1080/13803390600813841

Hofer, S. M., & Sliwinski, M. S. (2001). Understanding ageing. *Gerontology, 47,* 341–352, doi: 10.1159/000052825

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement. Special Issue: Test item banking, 10,* 369–380. doi: 10.1177/014662168601000405

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466. doi: 10.1037/0033-295X.104.3.427

Khaleefa, O., Abdelwahid, S. B., Abdulradi, F., & Lynn, R. (2008). The increase of intelligence in Sudan 1964-2006. *Personality and Individual Differences, 45,* 412–413. doi: 10.1016/j.paid.2008.05.016

Klauer, K. J. (1996). Teaching inductive reasoning: Some theory and three experimental studies. *Learning and instruction, 6,* 37–57. doi: 10.1016/S0959-4752(96)80003-X

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research, 78,* 85–123. doi: 10.3102/0034654307313402

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology, 27,* 1–25. doi: 10.1006/ceps.2001.1079

Lamiell, J. T. (2007). On sustaining critical discourse with mainstream personality investigators: Problems and Prospects. *Theory & Psychology, 17,* 169–185. doi: 10.1177/0959354307075041

Lamiell, J. T. (2011). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology.* Advance online publication. doi: 10.1016/j.newideapsych.2011.02.009

Lord, F. M., & Novick, M. N. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences, 11,* 273–285. doi: 10.1016/0191-8869(90)90241-I

Lynn, R., Hampson, S. L., & Millineux, J. C. (1987). A long-term increase in the fluid intelligence of English children. *Nature, 328,* 797. doi: 10.1038/328797a0

Lynn, R., Harvey, J. (2008). The decline of the world's IQ. *Intelligence, 36,* 112–120. Doi: 10.1016/j.intell.2007.03.004

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33,* 663–674. doi: 10.1016/j.intell.2005.03.004

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. In U. Neisser (Ed.). *The rising curve: Long-term gains in IQ and related measures* (pp. 183–206). American Psychological Association: Washington, DC, US.

Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for the Raven's Progressive Matrices. *Intelligence, 35,* 359–368. doi: 10.1016/j.intell.2006.10.001

Michell, J. (2011). Constructs, inferences, and mental measurement. *New Ideas in Psychology.* Advance online publication. doi:10.1016/j.newideapsych.2011.02.004

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review, 114,* 806–829. doi: 10.1037/0033-295X.114.3.806

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 699–710. doi: 10.1037/a0019182

Must, O., te Nijenhius, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence, 37,* 25–33. doi: 10.1016/j.intell.2008.05.002

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., …Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101. doi: 10.1037/0003-066X.51.2.77

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual information processing* (pp. 463–526). New York: Academic Press.

Olsen, W. & Morgan, J. (2005). A critical epistemology of analytical statistics: Addressing the sceptical realist. *Journal for the Theory of Social Behaviour,* 255–284. doi: 10.1111/j.1468-5914.2005.00279.x

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31,* 109–130.

Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology, 16,* 97–104.

Pirolli, P., & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12,* 235–275. doi: 10.1207/s1532690xci1203_2

Primi, R. (2002). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of Gf. *Intelligence, 30,* 41–70. doi: 10.1016/S0160-2896(01)00067-8

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods: Second edition.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., Congdon, R. T. & du Toit, M. (2011). *HLM 7: Linear and nonlinear modeling.* US: Scientific Software International.

Raven, J. C. (1938). *Progressive Matrices: A perceptual test of intelligence: Sets A, B, C, D, and E.* London, England: Lewis.

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II*. London: H. K. Lewis. (Distributed in the United States by The Psychological Corporation, San Antonio, TX).

Reder, L. M., Charney, D. H., & Morgan, K. I. (1986). The role of elaborations in learning a skill from an instructional text. *Memory & Cognition, 14,* 64–78.

Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26,* 337–356. doi: 10.1016/S0160-2896(99)00004-5

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435. doi: 10.1007/BF02306030

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path-mapping theory. *Cognitive Science: A Multidisciplinary Journal, 25,* 67–110. doi: 10.1016/S0364-0213(00)00035-5

Schaie, K. W. (2009). "When does age-related cognitive decline begin?" Salthouse again reifies the "cross-sectional fallacy." *Neurobiology of Aging, 30,* 528–529. doi: 10.1016/j.neurobiolaging.2008.12.012

Schiano, D. J., Cooper, L. A., Glaser, R., Zhang, H. C. (1989). Highs are to lows as experts are to novices: Individual differences in the representation and solution of standardized figural analogies. *Human Performance, 2,* 225–248. doi: 10.1207/s15327043hup0204_1

Schoenthaler, S. J., Amos, S. P., Eysenck, H. J., Peritz, E., & Yudkin, J. (1991). Controlled trial of vitamin—mineral supplementation: Effects on intelligence and performance. *Personality and Individual Differences, 12,* 351–362.

Sigman, M., & Whaley, S. E. (1998). The role of nutrition in the development of intelligence. In U. Neisser (Ed.). *The rising curve: Long-term gains in IQ and related measures* (pp. 155–182). American Psychological Association: Washington, DC, US.

Singley, K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational testing. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 263–331) New York: American Council for Education.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32,* 349−362. doi: 10.1016/j.intell.2004.06.004

Sundet, J. M., Borren, I., & Tambs, K. (2008). The Flynn effect is caused by changing fertility patterns. *Intelligence, 36,* 183–191. doi: 10.1016/j.intell.2007.04.002

Sundet, J. M., Eriksen, W., Borren, I., & Tambs, K. (2010). The Flynn effect in sibships: Investigating the role of age differences between siblings. *Intelligence, 38,* 38–44. doi: 10.1016/j.intell.2009.11.005

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: Simple mechanism to model of complex skill acquisition. *Human Factors, 45,* 61–76. doi: 10.1518/hfes.45.1.61.27224

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39,* 837–843. doi: 10.1016/j.paid.2005.01.0

Thorndike, E. L. (1922). *The elements of psychology.* New York, NY: A. G. Seiler.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlations between Operation span and Raven. *Intelligence, 33,* 67–81. doi: 10.1016/j.intell.2004.08.003

Van de Vijver, F. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678–709. doi: 10.1177/0022022197286003

Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118,* 339–356. doi: 10.1037/a0022749.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47,* 513–539. doi: 10.1146/annurev.psych.47.1.513

Vigneau, F., & Bors, D. A., (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65,* 109–123. doi: 10.1177/0013164404267286

Vigneau, F., & Bors, D. A., (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence, 36,* 702–710. doi: 10.1016/j.intell.2008.04.004

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34,* 261–272. doi: 10.1016/j.intell.2005.11.003

Viskontas, I. V., Holyoak, K. J., & Knowlton, B. J. (2005). Relational integration in older adults. *Thinking & Reasoning, 11,* 390–410. doi: 10.1080/13546780542000014

Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging, 19,* 581–591. doi: 10.1037/0882-7974.19.4.581

Wicherts, J. M., Boorsboom, D., & Dolan, C. V. (2010). Why national IQs do not support evolutionary theories of intelligence. *Personality and Individual Differences, 48,* 91–96. doi: 10.1016/j.paid.2009.05.028

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? *Intelligence, 32,* 509–537. doi: 10.1016/j.intell.2004.07.002

Williams, W. M. (1998). Are we raising smarter children today? School-and home-related influences on IQ. In U. Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 125–154). Washington, DC7 American Psychological Association.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 1047–1060. doi: 10.1037/0278-7393.15.6.1047

Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). *Psychological Review, 118,* 689–693. doi: 10.1037/a0024759

Woodworth, R. S., Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review, 8,* 247–261. doi: 10.1037/h0074898

Zelinski, E. M., & Kennison, R. F. (2007). Not your parents' test scores: Cohort reduces psychometric aging effects. *Psychology and Aging, 22,* 546–557. doi: 10.1037/0882-7974.22.3.546

BIOGRAPHICAL SKETCH

Mark C. Fox grew up in rural Michigan and has long been interested in understanding how people solve unfamiliar problems that require inductive or abductive reasoning.