Journal of Experimental Psychology: General
2013, Vol. 142, No. 3, 979–1000

0096-3445/13/$12.00 DOI: 10.1037/a0030155

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

Mark C. Fox
Florida State University

Ainsley L. Mitchum
SR Research, Kanata, Ontario, Canada

Secular gains in intelligence test scores have perplexed researchers since they were documented by Flynn (1984, 1987). Gains are most pronounced on abstract, so-called culture-free tests, prompting Flynn (2007) to attribute them to problem-solving skills availed by scientifically advanced cultures. We propose that recent-born individuals have adopted an approach to analogy that enables them to infer higher level relations requiring roles that are not intrinsic to the objects that constitute initial representations of items. This proposal is translated into item-specific predictions about differences between cohorts in pass rates and item-response patterns on the Raven's Matrices (Flynn, 1987), a seemingly culture-free test that registers the largest Flynn effect. Consistent with predictions, archival data reveal that individuals born around 1940 are less able to map objects at higher levels of relational abstraction than individuals born around 1990. Polytomous Rasch models verify predicted violations of measurement invariance, as raw scores are found to underestimate the number of analogical rules inferred by members of the earlier cohort relative to members of the later cohort who achieve the same overall score. The work provides a plausible cognitive account of the Flynn effect, furthers understanding of the cognition of matrix reasoning, and underscores the need to consider how test-takers select item responses.

Keywords: Flynn effect, Raven's Matrices, measurement invariance, cognitive aging, indeterminacy

Supplemental materials: http://dx.doi.org/10.1037/a0030155.supp

Intelligence test scores in developed nations rose dramatically during the 20th 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, Dolan, Carlson, & van der Maas, 2010). Contrary to intuition, the so-called Flynn effect is most pronounced on tests that were once regarded as culture-free such as Cattell's Nonverbal Intelligence Test (Lynn, Hampson, & Millineux, 1987) and the Raven's Matrices (Flynn, 1987). Culture in many countries has clearly changed since the early 20th century, 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 to dismiss environmental hypotheses altogether. Some have suggested that nutrition played a major role (Lynn, 1990; Sigman & Whaley, 1998), as there is evidence that nutritional supplementation can raise test scores (e.g., Schoenthaler, Amos, Eysenck, Peritz, & Yudkin, 1991). However, the effect sizes of nutritional supplementation are relatively small, and there is little regional or temporal correspondence between nutritional improvements and rising scores (Flynn, 1999). Mingroni (2007) suggested that the magnitude and stability of intelligence heritability estimates—heritabilities have remained stable 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.

By presuming that statistical patterns generalize across time periods and cultures, investigators often mistake local, relational characteristics of whole populations for universal, intrinsic properties of persons or items (e.g., Borsboom, Mellenbergh, & van Heerden, 2003; Lamiell, 2007; Wicherts & Johnson, 2009). Regardless of veracity, the biological hypotheses described above rest on a conceptual metaphor of the Flynn effect as an increase in some psychological *quantity* that is already possessed in greater or lesser amounts by every person in every population. Contrary to this interpretation, recent findings suggest the trend is better conceptualized as reflecting a *know-how* or *approach to problem solving*, a form of knowledge that proliferates only in relatively modern cultures. Item response models (Beaujean & Osterlind, 2008) and multigroup confirmatory factor analyses (Must, te Nijenhuis, Must, & van Vianen, 2009; Wicherts et al., 2004) reveal violations of mea-

This article was published Online First October 1, 2012.

Mark C. Fox, Department of Psychology, Florida State University; Ainsley L. Mitchum, SR Research, Kanata, Ontario, Canada.

The research was funded by National Institute of Aging Grant 3P01 AG17211. We thank Colleen M. Kelley, James R. Flynn, Joseph L. Rodgers, and Jelte M. Wicherts for commenting on a draft of this article, and we are thankful for additional input provided by Neil Charness, Walter R. Boot, Carol M. Connor, Anne E. Barrett, and Ralph Radach. We are also grateful to K. Anders Ericsson, whose influence on this article extends far beyond our references to his work.

Correspondence concerning this article should be addressed to Mark C. Fox, Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306-4301. E-mail: fox@psy.fsu.edu

surement invariance between cohorts, suggesting that the distributions of problem solving skills within a given region have changed over time. Other studies reveal that variation in fluid intelligence test scores diminished over time, particularly in the lower-performing half of the distribution (Colom, Lluis-Font, & Andrés-Pueyo, 2005; Teasdale & Owen, 2005). This implies that the proportion of very low performers declined more than the proportion of very high performers increased (Rodgers, 1998) and is compatible with the assumption that base-rate of individuals capable of accomplishing some necessary sub-goal of solving culture-free items has risen.

In this article, we seek to render specific, and therefore testable, the hypothesis that rising scores on the Raven's Matrices reflects adoption of an approach to analogy by recent-born individuals that enables them to infer higher-level relations requiring roles that are not intrinsic to the objects that constitute initial representations of items. This hypothesis has important implications for cross-cultural comparisons that we consider in the General Discussion.

Coping With Indeterminacy

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. An important implication of Flynn's proposal is that people have learned to search for and identify relations that are not immediately apparent given their initial interpretation of a problem. To show how this could improve performance on the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) Similarities—a test of acquired knowledge—Flynn and Weiss (2007, p. 217) considered the mental processes of a hypothetical child who supplies a "correct" similarity between *dusk* and *dawn*:

> You get up in the morning and go to bed at night but that makes no sense because I often sleep past dawn and go to bed after dark. They are alike in that the sky is half-lit and often very pretty but of course that is not always true. What they really have in common is that they are the beginning and end of both the day and the night. The right answer must be that they separate day and night.

Flynn and Weiss (2007) implied that children today are less prone than their grandparents to offering the first similarities (or dissimilarities) that they consider when comparing the two concepts. In what follows, we show that this account of how children have become better at diagnosing abstract relations need not be confined to transparent tests of acquired knowledge. The same basic idea can be translated into item-specific predictions about differences in patterns of item responses between cohorts on the seemingly culture-free Raven's Matrices.

Most people who are familiar with the analogs *solar system* and *atom* can solve the analogy, *sun is to planet as nucleus is to__* because the concepts *sun*, *planet*, and *nucleus* have familiar *roles* or functions that are intrinsic to their existence as concepts. To know a sun or a nucleus is to know that it attracts, and to know a planet is to know that it orbits.

Abstract items found on culture-free tests such as the Raven's Matrices are distinct from analogies like the one above in that appropriate responses call for higher-level relations requiring roles that are not intrinsic to the *objects*, or the "pieces" that constitute initial representations of items.[1] Many problem solvers would be stumped by the analogy, *&$B:B&$::T&T:$$_*, even if they are familiar with the symbols contained therein because no one's conception of *&* includes a role pertaining to *&*'s relation with *$* in this particular analogy; to know *&* or *$* is *not* to know that both are members of a pair.

The principal distinguishing feature of *&$B:B&$::T&T:$$_* then is not its unfamiliarity per se, but the *indeterminacy* of appropriate roles and relations with respect to how the problem is first represented (see Linhares, 2000, for some pictorial examples). Analogies like this one are difficult because objects themselves do not constitute knowledge about the roles or relations that characterize the analogy as a whole. Importantly, this does not imply that needed relations are complex (Carpenter, Just, & Shell, 1990) or even unfamiliar, but rather that they cannot simply be *read off of* the problem (e.g., Bunge, 1997; Chalmers, French, & Hofstadter, 1992; Linhares, 2000).[2] The *rule* or common relation needed to solve the analogy above is identical to the one that is needed to solve the most difficult Raven's Matrices items (Carpenter et al.'s, 1990, distribution-of-two-values rule) but is no less familiar than the principle used to sort one's socks.

Mapping Similar Objects

When objects in two or more analogs are similar, mapping can be accomplished by simply equating objects with their roles, and analogs with their relations. Such analogies epitomize concreteness, although they may still evoke impressions of "abstractness" if they call for little or no factual knowledge. For example, there is little problem of indeterminacy in 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 and the analogs) is self-evident precisely because objects are synonymous with their roles. Although this example may seem contrived, analogies in which objects serve as their own roles are common among easier items on tests like the Raven's Matrices.

Mapping Dissimilar Objects

A more flexible approach is needed to identify roles and relations when objects in two or more analogs are dissimilar (Chalmers et al., 1992), one that allows role to remain open like an

[1] It is beyond the scope of this article to discuss the important issue of how objects are defined in the first place. Our assumptions about which parts of items are objects to test-takers (and readers) should not be read as the claim that these objects exist *out there* independent of how problems are interpreted (see Chalmers et al., 1992). A finding that we do not discuss in the text is relevant: An analysis of think-aloud reports (Ericsson & Simon, 1980; Fox, Ericsson, & Best, 2011) collected from a small subset of Study 2 participants revealed considerable uniformity across persons and cohorts with respect to which potential objects were given verbal labels in Raven's Matrices items (e.g., "circle").

[2] The physicist and philosopher, Mario Bunge (1997, p. 420), has provided an excellent scientific exemplar of the argument made in this paragraph: "Astronomers can measure positions and velocities, but they cannot *read the law of gravitation off their data* [emphasis added]*:* such a law had to be *invented* [emphasis added] (and of course checked)."

unknown or a variable. Problem solvers can accomplish this by "acknowledging" that roles and relations are unknowns, and testing prospective roles and relations that defy initial interpretations of objects. This means actively searching for new roles, but more subtly, "understanding" that roles and relations are not necessarily compatible with initial representations.

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. For example, a very common relation such as *number of x* in *&$B:B&$::T&T:$$_* may be inferred quickly even if this relation is not sufficient by itself for mapping all of the objects (it does not apply to *T* and *B*). Immediately apparent roles and relations can be altered or combined if they do not enable mapping of relevant objects. For example, a modification may reveal that the more abstract relation, *two of a kind*, applies to every object in both analogs. This relation is more abstract because its role, *pair*, subsumes the concrete roles, *&*, *$*, *B*, and *T*, and returns the missing object, *&*. What distinguishes this approach from the superficial approach described above is that it allows roles to remain unknowns, even if only tentatively, until mapping is accomplished. We elaborate more on the distinction between concrete and abstract roles and relations by applying the same general principle to matrix reasoning.

Matrix Reasoning Tests

The items on all matrix reasoning tests are organized in a similar manner: Rules must be identified from the interrelations of objects in an array to determine which response choice would best complete the array. Our approach to identifying sources of item difficulty is influenced by Carpenter et al.'s (1990) taxonomy. However, we take into account how prospective objects are identified based on inferences about how participants within a given population characterize physical features of items. Each group of corresponding objects for a given item is classified according to the level of dissimilarity at which these objects must be mapped to infer a rule. This approach is distinct from rule taxonomies in that it replaces rigid generalizations about how people identify objects (i.e., operational definitions) with the flexible process of task analysis (e.g., Ericsson & Simon, 1993). Thus, one theory can accommodate two different populations even if members of these populations represent different features of the same items as objects, and one theory can be applied to tests comprised of different content. In this article, we compare populations that are similar enough to assume no difference in initial representations of objects.

Levels of Dissimilarity

Figure 1 is a relatively easy item that can be solved by mapping physically similar objects. Dots, each identical to the others, increase in number on the top from left to right. In addition, the number of dots decreases on the side from top to bottom. Notice 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 corresponding objects 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*.

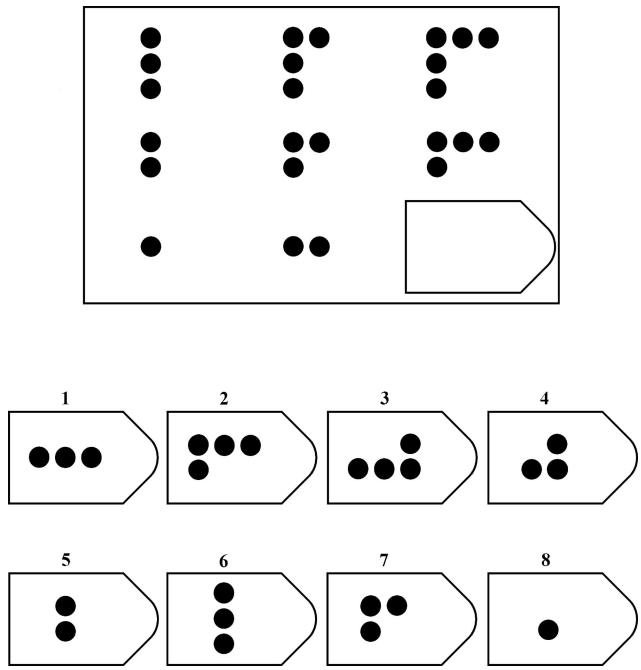

Figure 1. Identical dots decrease in number from top to bottom in columns and increase in number from left to right in rows. Both rules are classified as Level 1 because objects are synonymous with their roles (presence and placement within a figure). The answer is 1.

Both rules of Figure 1 occupy the lowest level of dissimilarity (Level 1) depicted in Table 1, 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 still yield a correct solution when applied to a rule at Level 1. At least one of the levels in the table is applicable to any rule of any item on the Raven's Matrices.

The levels are applied to Figures 2, 3, and 4 to illustrate the progression of increasing abstractness as a function of minor changes in features of items. Figure 2 is considered an addition-subtraction item in Carpenter et al.'s (1990) taxonomy 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*

Table 1
Level of Dissimilarity for Rules of 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 Raven's Matrices items.

middle (Level 2). Figure 3 is a simple modification of Figure 2 that requires a slightly more abstract version of addition or subtraction as it applies within single rows or columns: *one plus another equals the third* (the intermediate level in Table 1). By rearranging the objects in Figures 2 and 3, it is possible to create an item with the most abstract rule in Carpenter et al.'s taxonomy. Figure 4 is a distribution-of-two-values item, or as it is presented in Table 1, *two of a kind*. The role, *pair*, lacks similar physical appearance or placement in every row and column.

Figure 5, best illustrates the importance of the distinction between representation and actual physical features of an item. The three-sided shape in each row is clearly apprehended by the reader as an instance of the role, *triangle*, because people who have

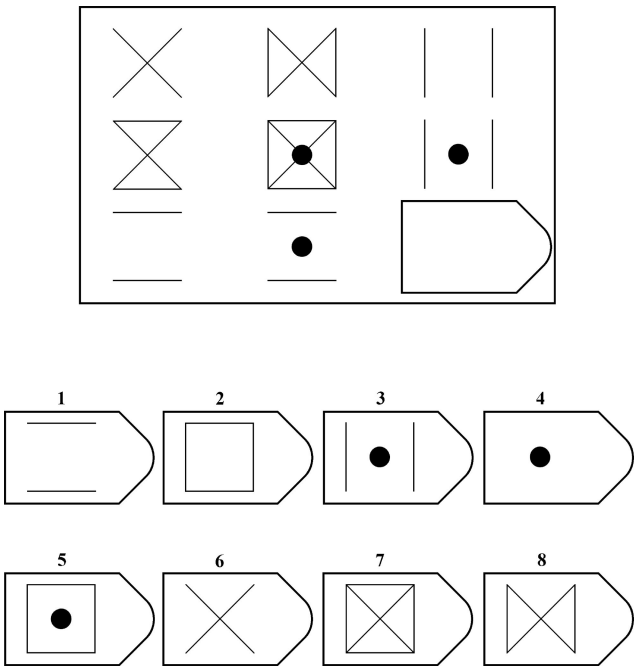

Figure 2. A figure-addition or subtraction and distribution-of-two-values item according to Carpenter et al.'s (1990) classifications. As an addition/subtraction item, Figure 2 is classified as Level 2 because objects with the same roles occupy the same location within rows or columns, but they do not necessarily appear similar because some objects are absent. The answer is 4.

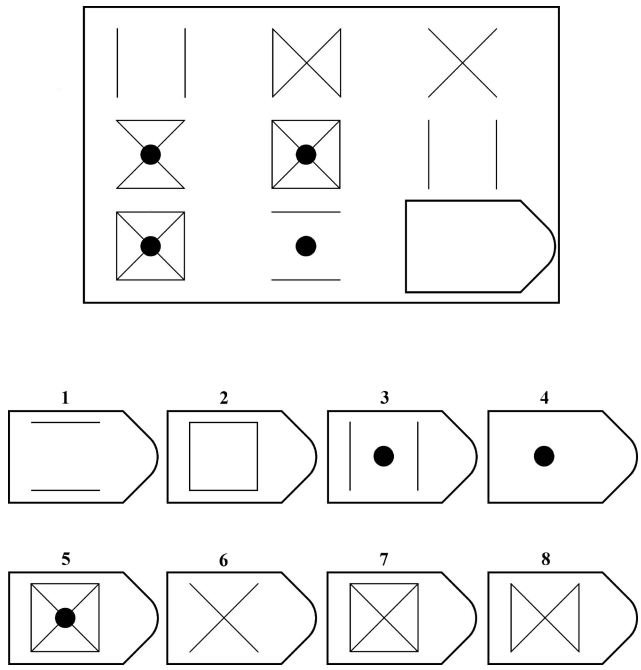

Figure 3. A modified version of Figure 2 that would be classified as the theoretical intermediate between Level 2 and Level 3 (see Table 1). Unlike Figure 2, Figure 3 cannot be solved with ordinary addition or subtraction (using whole rows or columns) because objects with the same role (e.g., *addend*, *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. The answer is 8.

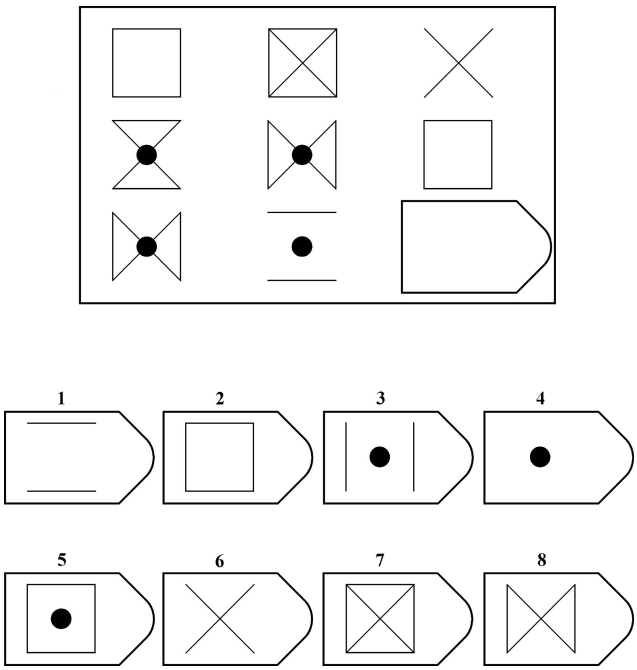

Figure 4. A modified version of Figures 2 and 3 that cannot be solved using an addition or subtraction rule. The item is classified as Level 3 because the abstract role, *pair*, does not have the same appearance or placement in every row or column. The answer is 7.

learned to read have also learned to regard three-sided shapes as members of the same category. However, these three-sided shapes are definitely not physically identical to one another. Because the role, *triangle*, may not be as universal as it seems, we commit to a formalist perspective for present purposes, whereby Figure 5 entails mapping objects that are dissimilar. In addition to differing in physical appearance, the corresponding objects 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). This example shows why no operational definition of dissimilarity can be expected to apply to every population unless there happen to be features of items that are perceived as objects by every person in every population (see Footnote 1).

As Table 1 shows, the notion of abstractness necessarily covaries with dissimilarity of objects because more abstract rules refer to features of items that differ from initial representations (Carpenter et al., 1990). 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 Raven's Matrices item can be solved by mapping objects at the third level of abstractness or lower.

Item Difficulty

It is possible that, contrary to what we have proposed, the level of dissimilarity at which objects must be mapped is not a source of item difficulty. Perhaps every person can map every object, regardless of level of dissimilarity, in much the same way that every person can use a pencil properly when selecting a response. If differences in level of dissimilarity elicit no within- or between-subjects variation, then it must be some other feature or features of items that make some more difficult than others.

It is also possible that level of dissimilarity is a source of difficulty in only one population and not another, or a major source of difficulty in one population and only a minor source in another. It is even possible that the same set of items elicits different rank-orders of difficulty for different individuals within a single population. Establishing that level of dissimilarity is a source of difficulty within at least one population is essential to establishing that the ability to map dissimilar objects varies between cohorts. The studies reviewed below report findings based on data collected mostly from American undergraduates born recently enough (most after 1970) to be considered samples of the same large, recent cohort.

There are two basic item dimensions that are known to moderate item difficulty: number of rules and dissimilarity of objects.

Number of rules. Studies reveal that items that require inferring more rules (referring specifically to number of tokens and not number of types throughout the article) to arrive at a solution are more difficult (e.g., Carpenter et al., 1990; Embretson, 1998). This finding is robust, at least within the population of young adults who have participated in matrix reasoning studies (Carpenter et al., 1990; Embretson, 1998; Primi, 2002).

Dissimilarity of objects. Consistent with the thesis of this article, there is also compelling evidence that the difficulty of items increases with the dissimilarity of corresponding objects.

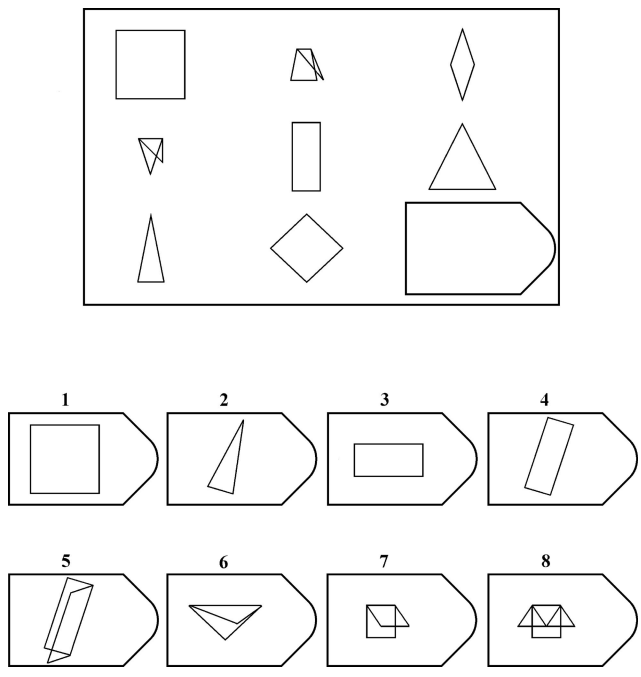

Figure 5. Corresponding objects are one of three shapes with wide and narrow versions, with or without a fold. The rules are classified as Level 3, but most formally educated problem solvers are likely to represent the shapes of a given type (e.g., triangle) as similar. The answer is 7.

Embretson (1998) and Primi (2002) constructed experimental items based on Carpenter et al.'s (1990) theory. Both studies found that items containing rules with dissimilar objects were more difficult than those containing rules with similar objects. Probabilities of solution were found to be lower when corresponding objects were not located in same rows and columns, and when corresponding objects were physically dissimilar to one another. Items generated by a program developed by Freund, Hofer, and Holling (2008) revealed the same pattern of findings as probabilities of solution were lower when the physical form of corresponding objects differed.

Based on our levels of dissimilarity, another study is also highly relevant. 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. Meo et al. found that the Raven's Matrices items were most difficult, followed by letter-like symbols, and then ordinary letters. This pattern of findings is compatible with our proposal. The corresponding objects of these new items are more similar than corresponding objects in the original test because the investigators essentially labeled the objects with either letters that are familiar to participants or with easily identified symbols; in other words, the investigators made the roles and objects identical to one another.

Consider the dissimilar objects in Figure 5. Such an item, based on Meo et al.'s (2007) classification, 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. From this same representational standpoint, two versions of the same letter-like symbol are somewhat less similar to one another than are ordinary letters even though both are physically very similar (albeit not necessarily identical) to one another because letter-like symbols are not as easily recognized as objects (i.e., they are not *chunks*). However, two versions of the same letter-like symbol are, to the extent that they can rightfully be regarded as the same symbol, more similar to one another than are any physically dissimilar objects in the Raven's Matrices.

Summary. The literature on matrix reasoning suggests two major sources of item difficulty within samples of individuals born after 1970. It is well-recognized that items with a greater number of rules are more difficult, and additional research suggests that items containing rules with dissimilar objects are also more difficult.

Hypothesis

No studies of matrix reasoning have examined sources of item difficulty in samples of younger adults from earlier cohorts. According to our analysis, individuals born more recently should find items containing rules with dissimilar objects easier to solve than did young adults born decades earlier. That is, if the Flynn effect reflects improvements in the ability to map dissimilar objects, then gains should be most pronounced on items with dissimilar objects.

Study 1: Predicting Changes in Item-Specific Pass Rate

The goal of Study 1 is to compare item-specific pass rates on the Raven's Matrices of two samples from two cohorts with virtually identical overall pass-rates. The first sample of pass rates was collected roughly four decades earlier than the second. We predict that higher pass rates in the more recent sample will be concentrated among items with rules containing dissimilar objects. This prediction of differences between cohorts in a specific skill, to be contrasted with differences in overall performance, is a prediction about measurement invariance, although we do not consider it in these terms until Study 2. It is further predicted that number of rules will correlate highly with pass rates within cohorts, but will not correlate with magnitude of changes in pass rate between cohorts.[3]

We consider one important caveat before proceeding. Item pass rates are not identical to item difficulties because they do not contain information about which individuals passed which items. Without this information, it is impossible to verify that the ordinal rank of difficulty for any two items within one cohort is uniform from person to person, that is, that difficulty is *distribution-free*. The need to satisfy this condition within the allowances of a probabilistic item response function is, of course, the basis of the Rasch model (e.g., Wright, 1977). From a strictly empirical standpoint, correspondence between predicted changes in pass rate and actual changes in pass rate would lend aggregate, *on-average* support to our proposal regardless of whether or not pass rates reflect distribution-free difficulties. However, the same findings would confer stronger, more nomothetic support for our proposal if pass rates do in fact reflect distribution-free difficulties because uniformity of difficulty implies greater generalizability of group findings to the individuals who comprise these groups. Accuracy data for Raven's Matrices items has been found to fit the Rasch model fairly well, at least compared to other tests that were not designed to meet this constraint (e.g., Gallini, 1983; Green & Kluever, 1992; van der Ven & Ellis, 2000; Vigneau & Bors, 2005). As Andrich (2004) recounts, the Raven's Matrices was one of the first tests found by Georg Rasch to fit his model. In Study 2, we use Rasch models to examine item difficulty for the same test with actual responses from comparable populations.

Method

Cohorts. Our goal was to locate sets of item-specific pass rates from at least two cohorts that are derived from large samples and separated by at least several decades. In conjunction with a standard literature search, we conducted a systematic search of

[3] As Wicherts and Johnson (2009) have shown, aggregate statistics such as heritabilities and differences between two populations in item-specific pass rate are expected to be greatest at pass rates of around 50% for a statistical reason that is logically distinct from any empirical hypothesis (i.e., 50% is the level of difficulty at which the most variance can be observed). Thus, hypotheses that predict a correlation between heritabilities and differences in pass rates will appear to receive empirical support regardless of whether they are true. This is not a problem in the present case because dissimilarity and number of rules, unlike heritabilities, are defined conceptually (from a psychological task analysis of items) rather than empirically (from an aggregate statistical analysis of items). Thus, correlations between differences in pass rate and dissimilarity or number of rules are not a logical certainty. Scatterplots confirm that neither dissimilarity nor number of rules evince the inverse U-shaped relationship with pass rate that would be expected if either variable were correlated with change in pass rate for the artifactual reason discussed by Wicherts and Johnson.

some 8,000 Raven's Matrices-related abstracts compiled by J. M. Wicherts for the terms *item analysis*, *item analyses*, *pass rate(s)*, and *proportion (in)correct*. We located nine articles that report item-specific pass rates or the information needed to calculate item-specific pass rates of non-clinical participants who completed the Raven Advanced Progressive Matrices Test: Arthur and Day (1994); Forbes (1964); Mitchum and Kelley (2010); Rushton, Skuy, and Bons (2004); Salthouse (1993); Unsworth and Engle (2005); Vigneau and Bors (2005); Wicherts and Bakker (2012); and Yates (1961).

Six of the nine studies—Forbes (1964), Mitchum and Kelley (2010), Rushton et al. (2004), Unsworth and Engle (2005), Vigneau and Bors (2005), and Wicherts and Bakker (2012)—can be divided into two comparable groups that are separated by about 50 years: Cohort 1940 and Cohort 1990. After discussing these two groups, we return to the remaining three studies.

Cohort 1940 consists of Forbes's (1964) young adults and late adolescents who were born around or shortly after 1940 and who were tested around 1961. Forbes's sample is, by itself, sufficiently large (n = 2,256) to provide reliable pass rates.

Cohort 1990 is a contemporary sample derived from combining the remaining data sets: Mitchum and Kelley (2010; n = 117), Rushton et al. (2004; n = 306), Unsworth and Engle (2005; n = 160), Vigneau and Bors (2005; n = 506), and Wicherts and Bakker (2012; n = 522). Original articles should be consulted for information about administration of the test, which varied across studies.[4] Birth years of the 1,611 participants span more than a decade, but item-specific pass rates are internally consistent (α = .98), and correlations of pass rates between any two samples exceed r = .90.

The two cohorts are closely matched on mean pass rate across items (Cohort 1940: M = .60, SD = .29; Cohort 1990: M = .63, SD = .27; d = 0.12). Overall pass rates in Forbes's (1964) sample are relatively high for the time period. The sample consisted of Air force recruits (n = 1,500), telephone engineering applicants (n = 500), and students at a teachers' training college (n = 256). Although Forbes expressed interest in discriminating at high levels of ability (pp. 223–224), he gave no indication that participants were sampled for high ability in particular. We cannot rule out the possibility that there is something unique about Forbes's participants that invalidates the present comparison between this sample and contemporary young adults. However, studies revealing violations of measurement invariance between cohorts (Must et al., 2009; Wicherts et al., 2004) are compatible with the assumption that the difference between Forbes's participants and their contemporaries is psychometrically distinct from the difference between Forbes's participants and modern test-takers.

Cohorts 1940 and 1990 were comparable in age at the time of testing with mean ages of about 20 years. Cohort 1940 is comprised of young adults and some late-adolescents. Cohort 1990 consists entirely of undergraduates from psychology department participant pools except for Rushton et al.'s (2004) sample, which consists of engineering students. Sex is confounded with cohort, as Cohort 1940 is primarily male (at least 66%), and Cohort 1990 is primarily female (roughly 60%–70%). There is evidence of a minor male advantage on the Raven's Matrices (Abad, Colom, Rebollo, & Escorial, 2004; Mackintosh & Bennett, 2005), but this advantage is not robust (Vigneau & Bors, 2008) and is probably too small to pose a concern given that the predicted advantage for Cohort 1990 on more abstract items will exceed the effect size of sex within cohorts if data are extreme enough to be interpreted as support for our proposal. Finally, Cohort 1940 data were collected in the United Kingdom, and Cohort 1990 data were collected in the United States, Canada, the Netherlands, and South Africa. Each region witnessed large Flynn effects (Flynn, 1987; te Nijenhuis, Murphy, & van Eeden, 2011).

Although we have provided an a priori basis for predicting that contemporary young adults (Cohort 1990) perform disproportionately better on items containing rules with dissimilar objects than their counterparts did 50 years ago (Cohort 1950), we cannot rule out the possibility that this hypothesis will be confirmed in the present study because of regional differences or unique effects of the specific periods from which these samples are drawn (see Rodgers, 1998), which may or may not be caused by the same factors responsible for the Flynn effect. In fact, our interpretation of the Flynn effect as a cohort effect rather than a series of distinct time period and/or region effects is a conceptual assumption that cannot be falsified by the study. It is noteworthy, however, that comparing each of the five data sets that constitute Cohort 1990 to Cohort 1940 in isolation reveals the same basic pattern of findings as those reported below. Interested readers can compare study-specific findings to the aggregate findings reported below by consulting Table 2.

Of the remaining three studies, Salthouse (1993) could not be used because pass rates of Items 23–36 are not reported. Yates's (1961) pass rates, derived from participants born around 1920, are too low to confer meaningful discriminations for the modern Advanced Progressive Matrices (Yates's, 1961, participants solved several easier items in addition to the 36 items that now constitute the Advanced Progressive Matrices), especially when comparing two distinct item variables. Only 11 of the 36 items have pass rates greater than 50%. Individuals who perform as low as participants in Yates's sample cannot be accommodated by our proposal without elaborating on the levels of dissimilarity in Table 1 through a task analysis of what are today considered very easy items (e.g., the Standard Progressive Matrices). Arthur and Day's (1994) data are comparable to Cohort 1940 and Cohort 1990 but were collected from a sample of participants who were born a decade too early to justify their inclusion in Cohort 1990. We do not include Arthur and Day's pass rates in Cohort 1990; however, Table 2 displays effect sizes of Arthur and Day's pass rates in relation to Cohort 1940. Overall item-specific pass-rates for cohorts 1940 and 1990 are displayed in Figure 6.

Item classifications. Carpenter et al.'s (1990, p. 431) classifications were used to assign number of rules to Raven's Matrices items. Carpenter et al. did not report numbers of rules for 11 of the 36 items either because the item could not be used 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, we assigned numbers of rules to the nine compatible items using Carpenter et al.'s taxonomy, and we as-

[4] For example, the test was administered with a 20-min time limit in Wicherts and Bakker (2012). An analysis including only those participants who completed nearly all of the items (34 of 36; n = 42) yielded results comparable to those reported in Table 2. Only 30 participants completed all 36 items.

Table 2
Study-Specific Effect Sizes for Cohort 1990 in Relation to Cohort 1940 as Partial Correlations With 95% Confidence Intervals

	Pass rate		Difference in pass rate from Cohort 1940	
	Dissimilarity	No. of rules	Dissimilarity	No. of rules
Study	Effect size [95% CI]	Effect size [95% CI]	Effect size [95% CI]	Effect size [95% CI]
Arthur & Day (1994)[a]	−.53 [−.73, −.24]	−.60 [−.78, −.34]	.43 [.12, .66]	.10 [−.24, .41]
Mitchum & Kelley (2010)	−.46 [−.69, −.15]	−.60 [−.78, −.34]	.33 [.02, .58]	−.06 [−.38, .27]
Rushton et al. (2004)	−.47 [−.69, −.17]	−.57 [−.76, −.30]	.55 [.27, .74]	.19 [−.15, .49]
Unsworth & Engle (2005)	−.56 [−.75, −.28]	−.60 [−.78, −.34]	.34 [.01, .60]	.08 [−.26, .40]
Vigneau & Bors (2005)	−.52 [−.72, −.23]	−.57 [−.76, −.30]	.43 [.12, .66]	−.08 [−.40, .26]
Wicherts & Bakker (2012)	−.45 [−.68, −.14]	−.65 [−.81, −.41]	.55 [.27, .74]	−.38 [−.63, −.06]
Overall	−.50 [−.71, −.21]	−.61 [−.78, −.35]	.60 [.33, .78]	−.10 [−.41, .24]

Note. Partial correlations reflect unique variance shared between item variables and pass rate or difference in pass rate in relation to Cohort 1940.
[a] Study is not included in Cohort 1990 because data were collected from test-takers who were born too early to be in this cohort.

signed numbers of rules to the remaining two items using novel rules. These decisions did not alter the pattern of results reported below. Interested readers can compare study-specific findings to the aggregate findings reported below by consulting Table 2 (study-specific pass rates are available in the supplemental materials along with all of the data analyzed in this article).

Classifying rules according to the dissimilarity of corresponding objects requires nothing less than knowledge of how participants represent objects. Because this information is unavailable, simple criteria were used to optimize simplicity and plausibility including the assumption that members of Cohorts 1940 and 1990 have the same objects. In a manner consistent with classifications in studies reviewed above, we defined similarity with respect to the positions and physical features of the parts of items that occupy the same role based on relations that are compatible with correct answers. Given these criteria, corresponding objects that differ in size, but remain otherwise identical (this includes lengths of single lines; e.g., Item 10), and shading patterns (which may appear on different shapes; e.g., Item 21) are considered similar. As noted, physically distinct shapes like the three triangles in Figure 5 are considered dissimilar.

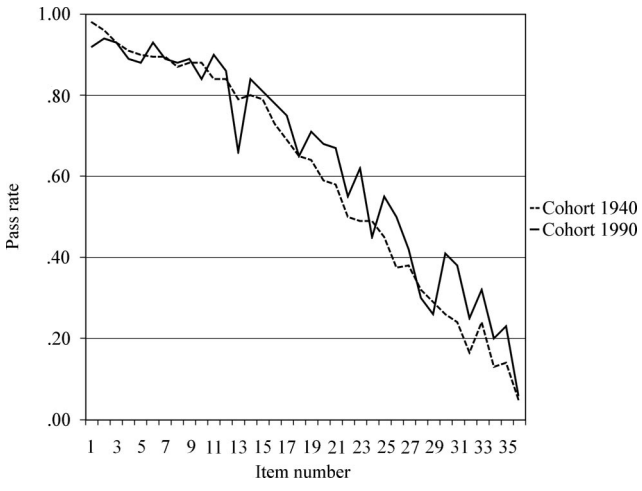

Figure 6. Comparison of item-specific pass rates of Cohorts 1940 and 1990 for Study 1.

Each rule of every item was classified as one of the three levels of dissimilarity presented in Table 1. Rules were assigned to the lowest (most similar) level that is sufficient for correct mapping of objects. Because our proposal makes no assumptions about whether participants represent columns or rows as analogs, the lowest level of dissimilarity compatible with solution was established by comparing objects from row to row and column to column. In accord with Table 1, 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.

The Appendix shows classifications at the level of individual rules. Not surprisingly, these classifications overlap considerably (about $r = .6$) with a variable representing Carpenter et al.'s (1990) ranking of rules by difficulty.

Results

The limited number of observations (one for each of 36 items per sample) preclude sophisticated regression models for examining relationships between level of dissimilarity, number of rules, and the outcome variables of pass rate and change in pass rate. However, because number of rules and dissimilarity are correlated ($r = .42$), linear regression is used to obtain partial correlations representing unique variance shared between either predictor variable and pass rates or change in pass rates.

It was found that number of rules and level of dissimilarity both correlate with item-specific pass rates in both cohorts. In regression models, the two variables account for about two thirds of the variance in pass rate in Cohort 1940 ($R^2 = .66$) and Cohort 1990 ($R^2 = .61$). Effect sizes for number of rules are large in Cohort 1940 ($r = –.68$, 95% CI [–.82, –.45]) and Cohort 1990 ($r = –.70$, 95% CI [–.84, –.48]), as pass rate was found to decrease with greater numbers of rules. The effect sizes for level of dissimilarity are comparable to those for number of rules in Cohort 1940 ($r = –.69$, 95% CI [–.83, –.47]) and Cohort 1990 ($r = –.62$, 95% CI [–.79, –.37]). These results are consistent with findings of Carpenter et al. (1990), Embretson (1998), and Primi (2002).

Because number of rules correlates with level of 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 1940: $r = –.59$, 95% CI [–.77, –.32]; Cohort 1990: $r = –.61$, 95% CI [.35, .78]) and level of dissimilarity (Cohort 1940: $r = –.61$, 95% CI [–.78, –.35]; Cohort 1990: $r = –.50$, 95% CI [–.71, –.21]) are strong independent predictors of pass rate in both cohorts. These within-cohort results concur with previous research and support the hypothesis 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 1940 to Cohort 1990 were calculated by subtracting item-specific pass rates of Cohort 1940 from those of Cohort 1990. These changes in item-specific pass rates approximate a normal distribution (kurtosis and skewness are within the range of ± 1; Kolmogorov–Smirnov $Z = .42$) and are treated as a continuous dependent variable in the following analysis.

Regression revealed a small effect size for number of rules ($r = .20$, 95% CI [–.14, .50]) that remains small when unique variance is isolated ($r = –.10$, 95% CI [–.41, .24]). The effect size of dissimilarity is larger ($r = .61$, 95% CI [.35, .78]), but it too remains virtually unchanged when unique variance is isolated ($r = .60$, 95% CI [.34, .78]). This confirms our prediction that recent-born individuals outperform their predecessors primarily on items that require mapping dissimilar objects.

Figure 7 is a scatterplot of differences in pass rates between Cohorts 1940 and 1990 as a function of level of dissimilarity. As predicted, level of dissimilarity is positively associated with changes in pass rates, as Cohort 1990 gains were concentrated in items with dissimilar corresponding objects. Although number of rules may also be associated with change in pass rate, the association appears to be a consequence of items with more rules also tending to have rules with dissimilar objects.

Table 2 confirms that the same pattern of findings is obtained by comparing any individual Cohort 1990 group to Cohort 1940.

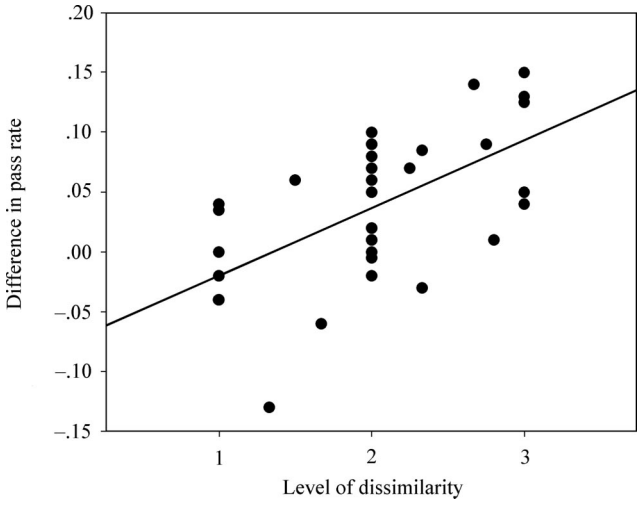

Figure 7. Differences in pass rates between Cohorts 1940 and 1990 as a function of level of dissimilarity. Differences reflect Cohort 1990 pass rates minus Cohort 1940 pass rates.

Discussion

Study 1 shows that number of rules and level of dissimilarity are sources of variation in item-specific pass rates from two large samples collected nearly 50 years apart.

As predicted, cohort-related gains in pass rates are associated with level of dissimilarity, but not number of rules. These item-specific gains in pass rate on the test with the largest Flynn effect are consistent with the assumption that young adults became better at mapping dissimilar objects over time. The comparison revealed the expected difference in pass rates for items with rules containing dissimilar objects despite equivalence in overall scores of the two cohorts. The effect size of $r = .60$ is fairly large despite being constrained by a correlation between the variables representing level of dissimilarity and number of rules. Although results of Study 1 should not be taken for granted to generalize across all between-cohort comparisons, Table 2 suggests that they are robust enough to generalize across several distinct populations. Even by itself, Study 1 offers compelling evidence that rising scores reflect changes in the means by which people map dissimilar objects. However, it is also important to consider that differences between pass rates of items are not identical to differences between difficulties of items.

It is illustrative at this point to consider our Study 1 prediction from the perspective of measurement invariance (e.g., Millsap, 2007), in particular, measurement invariance as it relates to achievement of actual goals that must be accomplished to select a correct response (without guessing). In matrix reasoning, these goals are mapping objects at one or more levels of dissimilarity for each of one or more rules. By proposing that members of Cohort 1990 map objects at higher levels of dissimilarity than Cohort 1940 participants who achieved the same overall pass rates, we also predicted that Raven's Matrices scores either overestimate the level of dissimilarity at which Cohort 1940 participants map objects relative to Cohort 1990 participants, or underestimate the number of rules that Cohort 1940 participants infer relative to Cohort 1990 participants (or some combination of both). In other words, members of Cohort 1940 either map objects at lower levels of dissimilarity than Cohort 1990 participants who achieve the same score on the Raven's matrices, or infer a greater number of total rules than Cohort 1990 participants who achieve the same score (or some combination of both).

Although Study 1 yielded findings consistent with this prediction, we were unable to verify the prediction conclusively because Study 1 data consisted of pass rates rather than difficulties. We cannot regenerate true difficulties out of 50-year-old pass rates, but we can test the same prediction in a contemporary cross-sectional sample consisting of present-day younger adults and older adults, the latter of whom were roughly the same age as Forbes's (1964) participants at the time that he collected his data.

Study 2: Measurement Invariance

Study 2 tests the prediction that Raven's Matrices scores violate measurement invariance in relation to response categories defined by goals that must be achieved to generate correct responses.

Raven's Matrices responses were obtained from a cross-sectional sample of participants from two cohorts (i.e., younger and older adults) separated by roughly 50 years.[5] Although the use of contemporary younger and older adults is not ideal, 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 prospective cause that motivated predictions. Second, findings verifying our prediction that younger adults will map objects more successfully at higher levels of dissimilarity cannot be attributed to biological aging without providing an alternative explanation of the primary finding of Study 1, namely, 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 for measurement invariance is much like testing for any other interaction. According to Millsap's (2007) definition, a test (e.g., the Raven's Matrices) is measurement invariant in relation to a group variable (e.g., cohort) and a criterion variable "if and only if" (p. 463) the probability of achieving a score on the test given group membership and placement along a criterion variable is identical to the probability of achieving the same score given only placement along the criterion variable.

The criterion variables for present purposes are the actual problem solving goals achieved by participants as indicated by the correspondence between actual item responses and correct item responses. More specifically, criterion variables are defined by the level of dissimilarity at which participants were able to map objects and the number of rules they were able to infer according to features of their actual item responses. From this perspective, the hypothesis of this article is that Raven's Matrices scores violate measurement invariance, either by overestimating the level of dissimilarity at which members of earlier cohorts map objects relative to members of later cohorts, or underestimating the number of rules that members of earlier cohorts are able to infer relative to members of later cohorts. Verifying this prediction would help to justify our earlier conclusions by revealing that there is nothing paradoxical about Cohort 1940 participants in Study 1 achieving the same overall pass rates as Cohort 1990 participants despite having lower pass rates on items with dissimilar corresponding objects.

Polytomous Rasch Models

Polytomous models distinguish between response choices within single items, making it possible to define latent variables representing the level of dissimilarity at which participants map objects, and the number of rules that participants infer, by classifying the response choices for every item. Thus, it is possible to express our predictions within the confines of a preexisting test such as the Raven's Matrices by creating latent variables that are distinct from raw score or accuracy (e.g., Kelderman, 1996).

Masters's (1982) partial credit model (PCM) is the foundation of the models presented below. The PCM combines the unique conceptual properties of the Rasch model (Wright, 1977) with the allowance of multiple response categories. Both the Rasch model and its PCM extension posit distribution-free scaling; that is, knowledge of one person's ability, *as defined within the confines of the model*, is fully contained within his or her responses, and is not furthered in any way by comparing his or her responses to those of others (e.g., Wright, 1977). However, unlike the dichotomous Rasch model, which accommodates only accuracy data, the polytomous PCM allows levels of a latent variable to have an ordering that is distinct from the ordering of number of correct responses. That is, one participant can score higher on the test than another participant in terms of number of correct responses, but still place lower on the latent variable. This means that the PCM can accommodate the prediction that Cohort 1940 participants place higher on a latent variable than Cohort 1990 participants who achieve a lower raw score on the Raven's Matrices, or conversely, that Cohort 1940 participants place lower on a latent variable than Cohort 1990 participants who achieve a higher raw score.

The PCM transposes the dichotomy of the Bernoulli distribution from accuracy of item response to the probability of responding in Category k relative to one or more additional ordinal categories. The probability of responding in Category k of item i is

$$P_{ik}(\theta) = \frac{\exp\sum_{j=0}^{k}(\theta - \delta_{ik})}{\sum_{i=0}^{m-1}\exp\sum_{j=0}^{k}(\theta - \delta_{ik})},$$

where δ_{ik} is the difficulty parameter, and θ is the latent variable corresponding to the ability expressed by the model. Figure 8 illustrates the model with hypothetical item category response functions. The predicted probability of a response in, say, Category 3 relative to Category 2 increases with higher placement along the latent variable. Considering only the two most extreme values of θ (the far left and the far right of the graph), it is clear that individuals low in ability are almost as likely to respond in Category 3 as Category 2, whereas individuals high in ability are much more likely to respond in Category 3 than Category 2.

We utilize two polytomous models. The *dissimilarity model* equates ability with the level of dissimilarity at which participants can successfully map corresponding objects. Responses containing no correct objects or correct objects for rules with similar corresponding objects occupy lower categories, and responses containing correct objects for rules with dissimilar corresponding objects in addition to correct objects for rules with similar corresponding objects occupy higher categories. In other words, the model takes for granted that participants who map objects for, say, a Level 3

[5] A unique study by Babcock (2002) has revealed similar response patterns for contemporary older and younger adults with respect to the kinds of errors they commit according to Forbes's (1964) taxonomy of errors. It is unclear whether Babcock's results are compatible with our own. If older adults have particular difficulty with mapping dissimilar objects, they should be more likely than younger adults to make wrong-principle and repetition errors, and possibly confluence-of-ideas errors (see Forbes, 1964), because these error types suggest failure to infer rules. Although the finding was not deemed significant by conventional standards, Babcock did find that a greater proportion of older adult errors were wrong-principle errors, while repetition and confluence-of-ideas errors were more similar across groups. An important caveat of Babcock's study is that patterns of responses and error rates for more difficult items may have been systematically biased if the very brief time limit of 20 min (half of the optional standardized time limit) forced many participants to guess at these items or forego providing a response.

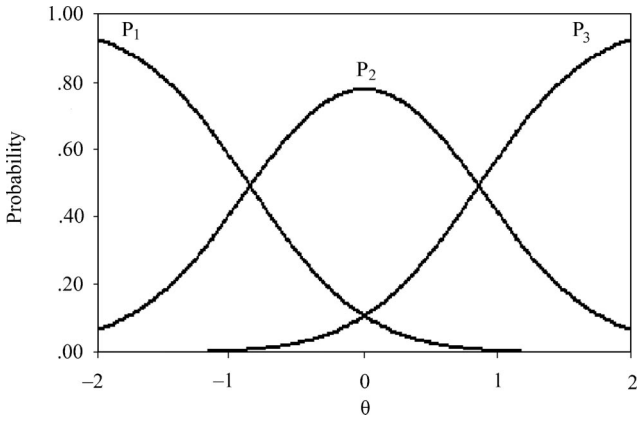

Figure 8. Partial credit model item category response functions for a three-category item. The predicted probability of a response in Category 3 relative to Category 2 increases with higher placement along the latent variable. Considering only the two most extreme values of θ (the far left and the far right of the graph), the model assumes that individuals with low ability scores are almost as likely to respond in Category 3 as Category 2, whereas individuals with high ability scores are far more likely to respond in Category 3 than Category 2.

rule, can and do map objects for a Level 2 rule within the same item.

In contrast, the *number-of-rules model* equates ability with the number of rules inferred by a participant by assuming that high ability participants select responses with more correct objects. The number-of-rules model places responses containing correct objects for, say, two rules, in a higher category than responses containing correct objects for only one rule, regardless of similarity of objects. Both models place responses with no correct objects in the lowest possible category. Notice that ordinal categories themselves are distinct from individual goals or "steps" (Masters, 1982, p. 155) within these categories (viz., rules).

Predictions and assumptions of the two models coincide with one another to the extent that response options containing correct objects at higher levels of dissimilarity also contain correct objects at lower levels of dissimilarity (i.e., to the extent that some response options are simply closer to being correct than others based on our criteria). Response options for both variables are not perfectly counterbalanced within items and there is linear dependence between the two variables both within and between items (i.e., correlations between response and item vectors representing levels of variables). For these reasons, we cannot test a multidimensional model that includes both variables (e.g., see Adams, Wilson, & Wang, 1997). However, we emphasize that our primary goal is to test predicted violations of measurement invariance on a preexisting test with a large Flynn effect rather than to advance a definitive model of matrix reasoning.

As a baseline reference for the polytomous models, and a means of testing our earlier assumption that aggregate pass rates of Raven's Matrices items can be interpreted as if item difficulties are distribution-free, we also test a dichotomous model that accommodates only accuracy data. This model equates ability with success at solving Raven's Matrices items rather than any claim about how item responses are selected.

Testing for Violations of Measurement Invariance

In the context of individual tests, measurement invariance is generally evaluated at the level of individual items. An item exhibits differential item functioning (DIF; sometimes called item bias) to the extent that members of one group who respond in the same category as members of the other group achieve a higher or lower raw score on the test.[6]

Miller and Spray's (1993) logistic discriminant function analysis is used to detect DIF because it is applicable to polytomous items and is more powerful than parametric approaches and nonparametric alternatives such as the generalized Mantel–Haenszel procedure (Miller & Spray, 1993; Su & Wang, 2005) and multinomial logistic regression (Hidalgo & Gomez, 2006). The procedure does not appear to inflate Type I error relative to these alternatives (Hidalgo & Gomez, 2006; Su & Wang, 2005). Logistic discriminant function analysis is applied within a logistic regression framework by interchanging conditional and fixed variables such that group membership is conditioned on raw score and response category. Thus, cohort membership becomes the dependent variable, and detection of DIF becomes a matter of determining whether or not response category improves predictions of cohort membership above and beyond overall raw score on the Raven's Matrices. To the extent that it does, Raven's Matrices scores violate measurement invariance in relation to cohort and latent variables as defined by response categories.

Method

Items and participants. Because the Flynn effect is, first and foremost, an effect of raw scores, the most valid indicator of Raven's Matrices score is number correct. Thus, our analysis is limited to complete sets of responses to a common set of items. The sample consists of 260 older (Cohort 1940) and younger (Cohort 1990) participants. Each participant completed a computerized version of the Raven's Matrices within a 3-year period spanning from 2008 to 2010. This includes 50 participants from Boot et al. (2012). Boot et al. omitted four items from their version of the test (Items 21, 25, 29, and 33). These items are excluded from the analysis in keeping with the criteria specified above. The final data set consists of 32 items completed by 223 participants (Cohort 1940: $n = 72$, mean age = 73 years; Cohort 1990: $n = 151$, mean age = 19 years). Cohort 1940 participants in the present study were born around the same time as Cohort 1940 participants in Study 1.

Response classifications. The eight response options for each of the 36 Raven's Matrices items (one correct response and seven lures) were categorized according to the dissimilarity and number-of-rules models using the same criteria as used in Study 1.

In the dissimilarity model, responses were categorized according to the level of dissimilarity of correct objects such that each

[6] Both our predictions and our use of "DIF" in the text refer to *uniform* DIF in particular. Psychometricians often distinguish between uniform DIF, or bias that is constant across levels of ability, and *non-uniform* DIF, or bias that interacts with level of ability (e.g., Su & Wang, 2005). Because we neither predicted nor found substantial non-uniform DIF, we forgo discussing it in the text to evade a potentially confusing distinction that is incidental to our thesis. Statistics for non-uniform DIF have been made available in Table 3 for interested readers.

response was placed in the category corresponding to the lowest level of dissimilarity for any of its correct objects. For example, the eight choices for an item with one Level 2 rule and one Level 3 rule contain either (1) no correct objects, (2) the correct object for the Level 2 rule and incorrect object for the Level 3 rule, (3) the incorrect object for the Level 2 rule and correct object for the Level 3 rule, or (4) correct objects for both the Level 2 and Level 3 rules. The ordinal categories for these choices are 1, 2, 1, and 3, respectively. If there is only one Level 3 rule, it follows that Category 3 admits only one response, the correct response. Other responses that contain a correct object for the Level 3 rule necessarily contain incorrect objects for the Level 2 rule and are categorized as 1.

In the number-of-rules model, response choices were categorized according to the number of correct objects that they contain. Response options for an item with two rules contain either no correct objects, one correct object, or two correct objects. Thus, the ordinal categories for these responses are 1, 2, and 3, respectively. Response categories for both models are shown in the Appendix.

The imperfect correspondence between the hypothesis and Raven's Matrices response categories leads to several limitations of the models. The deviation of the dissimilarity model from the dichotomous model (i.e., accuracy data) is limited because many items containing rules with Level 2 or Level 3 dissimilarity do not contain partially correct responses (see the Appendix). This makes it difficult to confirm the prediction that Raven's Matrices scores overestimate the level at which Cohort 1940 participants map objects.

In addition, some incorrect responses of some items contain odd variations of objects that are incompatible with rules as defined by Carpenter et al. (1990) and the criteria used in the first two studies. For example, each figure of the matrix in Item 14 contains the same invariant object (a "Y" shape rotated 90° clockwise). As an invariant "constant," this object is exempt from Carpenter et al.'s (1990) rules but must be present in the correct answer. In fact, the object is present in every response option, but is incorrectly inverted in one option that would otherwise be considered correct because it contains correct objects for both of the two rules. Incorrect response options like this one cannot be categorized in accordance with either model without making additional assumptions about how participants solve items. However, they cannot be excluded without reintroducing the problem of missing data. Our solution was to place the 18 response options for nine items like this one into the lowest response categories. Although not strictly consistent with either model, this can only decrease confirmation of our predictions because it reduces the degree to which scores on the polytomous variables can deviate from ordinary raw scores.

Differential item functioning. Logistic discriminant function analysis (Miller & Spray, 1993) simplifies the otherwise awkward application of logistic regression to polytomous items by exchanging the categorical predictor (group) and binary dependent variable (accuracy) such that the regression represents the conditional probability of membership in one group versus the other given raw score (number of Raven's Matrices items answered correctly), the criterion variable (response category), and an interaction term for non-uniform DIF (Raven's Matrices score by response category; see Footnote 6). The p values of changes in chi-square in the stepwise procedure are the probability of obtaining data at least as extreme as those observed if there is no DIF to be found at the level of the population (see Miller & Spray, 1993, for more details). The procedure is applied to every item in isolation.

Results

The analysis can be decomposed into the two basic stages of first evaluating the fit of the models to a population comprised of Cohort 1940 and Cohort 1990 participants, and then testing Raven's Matrices scores for measurement invariance in relation to cohort and item-level response categories as defined by the models (i.e., DIF).

Accuracy and fit. Cohort 1990 participants achieved higher raw Raven's Matrices scores than Cohort 1940 participants (Cohort 1990: $M = 17.11$, $SD = 5.11$, 95% CI [16.29, 17.93]; Cohort 1940: $M = 10.79$, $SD = 5.01$, 95% CI [9.63, 11.94]; $d = 1.25$), and they achieved higher sum-scores on dissimilarity (Cohort 1990: $M = 22.70$, $SD = 6.27$, 95% CI [21.60, 23.80]; Cohort 1940: $M = 16.00$, $SD = 5.81$, 95% CI [14.70, 17.30]; $d = 1.11$) and number-of-rules (Cohort 1990: $M = 43.60$, $SD = 9.07$, 95% CI [42.20, 45.00]; Cohort 1940: $M = 35.10$, $SD = 8.23$, 95% CI [33.2, 37.0]; $d = 0.98$) variables. The between-cohort effect size for Raven's Matrices scores ($d = 1.25$) is fairly representative of both the Flynn effect and typical findings in cross-sectional studies of cognitive aging.

In accord with distribution-free assumptions of the Rasch model, conditional maximum likelihood was used to estimate item and person parameters independently. Because the overall fit of the data to all three models (dissimilarity, number-of-rules, and dichotomous models) is very good, we present more specific item-level fit statistics in the form of weighted mean-squares (Wright & Masters, 1982). Values of 1 indicate ideal fit, values of less than 1 indicate less variation than predicted by the model, and values of greater than 1 indicate greater variation than predicted (unexplained variation). Interpretation of a given value is unaffected by number of participants (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008), in part, because the denominator is the degrees of freedom.

Fit statistics for all three models are displayed in Figure 9. As the figure shows, patterns of responses are fairly compatible with

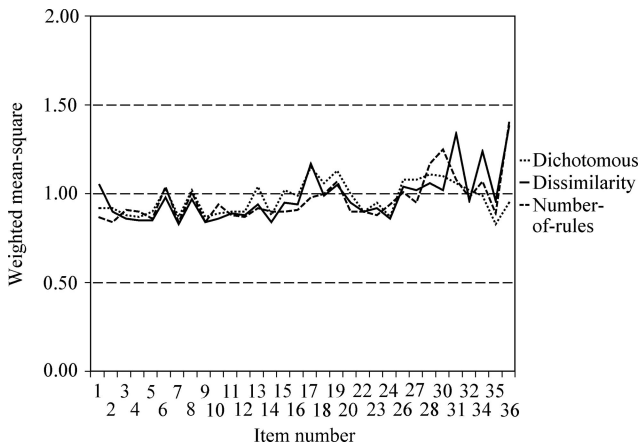

Figure 9. Weighted mean-square by item for all three models. The solid horizontal line represents ideal fit, and the dotted horizontal lines represent rule-of-thumb boundaries for acceptable fit.

all three models. Excellent fit of responses to the dichotomous model lends justification to our interpretation of Study 1 pass rates as indicators of item difficulty. Correspondence of data with the dissimilarity model suggests that the level of dissimilarity at which participants map objects is distribution-free, but we reiterate that the Raven's Matrices does not permit the dissimilarity variable to deviate far enough from raw score to justify any firm conclusions (see the Appendix). Finally, response patterns are also compatible with the number-of-rules model, which shows that difficulties are distribution-free when difficulty is defined by number of rules in items within these populations.

Fit of the data to the polytomous models is compatible with the interpretation of dissimilarity and number of rules as distribution-free sources of item difficulty. The pivotal question is whether the relationship between ordinary Raven's Matrices scores and the problem goals corresponding to scores on these variables are comparable in Cohorts 1940 and 1990.

Differential item functioning. DIF statistics for Raven's Matrices scores in relation to response categories, derived with Miller and Spray's (1993) procedure, are presented in Table 3. Raven's Matrices scores exhibit at least some degree of DIF in relation to variables defined by each of the three models. The magnitude of changes in chi-square reveal that the number-of-rules variable manifests far greater DIF for a greater number of items ($n = 18$) than the dissimilarity variable ($n = 8$) or the dichotomous variable ($n = 5$) using the arbitrary criterion of $p < .05$. In fact, the number-of-rules variable reveals substantial DIF for every item that shows any DIF for the other variables. This suggests that Raven's Matrices scores violate measurement invariance between cohorts by either overestimating or underestimating the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants.

The directionality of DIF for the number-of-rules variable and dissimilarity variable is ascertained by evaluating response patterns within items.

Number-of-rules. An example helps to illustrate how directionality of DIF was assessed for the number-of-rules variable. Figure 10 shows a logistic discriminant function analysis (see Miller & Spray, 1993) for Item 28. The *y*-axis represents the log-odds of Cohort 1990 membership when cohort membership is

Table 3
Differential Item Functioning as Indicated by Raven's Matrices Scores in Relation to Cohort and Response Categories

	Dichotomous (accuracy)		Dissimilarity		Number-of-rules	
Item	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform
1	1.61	0.27	1.88	1.15	3.00*	0.26
2	2.65**	0.59	2.65**	0.59	5.06**	0.35**
3	8.38**	2.13	8.38**	2.13	8.99**	7.09**
4	0.42	0.15	0.42	0.15	5.32*	0.20*
5	2.16	0.51	2.16	0.51	4.09*	4.10*
6	0.35**	0.32	0.35**	0.32	14.71**	0.83
7	6.98*	0.04	6.98*	0.04	6.98*	0.04
8	0.59	0.13	0.59	0.13	3.40	2.96
9	0.08	0.02	0.07	0.02	0.08*	0.02
10	0.34	2.40	0.34	2.40	6.11*	2.83
11	0.34	0.48	0.34	0.48	0.34	0.48
12	1.30	1.22	1.30	1.22	1.30	1.22
13	0.13*	2.11	2.88*	2.68	3.05*	2.67
14	5.98*	0.04	5.98*	0.04	6.30*	1.84*
15	3.60	0.88	3.60	0.88	6.12*	4.57
16	0.00	0.60	0.00*	0.60	0.00*	0.60
17	1.93**	0.90	4.31**	3.60	4.31**	3.60*
18	9.93**	0.05	9.93**	0.05	10.57**	4.60*
19	0.01	0.15	0.01	0.15	4.33*	0.20
20	0.49	0.00	0.49	0.00	0.49	0.00*
22	0.02	0.02	0.02	0.02	0.69	6.30*
23	0.10	1.79	0.10	1.79	0.10	3.37
24	0.02	0.04	0.02	0.04	1.01	3.37
26	0.23	0.01	0.23	0.01	0.44	1.93
27	0.03	0.06	0.03	0.06	2.75**	0.21
28	3.44*	0.73**	3.44*	0.73**	10.29**	2.10**
30	4.15*	9.36**	4.15*	9.36**	8.10**	18.91**
31	0.25	2.60	0.12	2.06	5.68**	6.68**
32	0.64	2.09	1.69*	2.84	9.06**	2.65
34	3.80	0.01	4.25*	2.03	4.25*	2.03
35	1.50	0.02	1.15**	0.76*	1.51*	0.76*
36	0.15	3.59	7.81**	4.13*	3.93*	5.96*

Note. Uniform and non-uniform differential item functioning (DIF) as indicated by raw Raven's Matrices score in relation to cohort and response category for dichotomous, dissimilarity, and number-of-rules variables. The text refers only to uniform DIF (see Footnote 6). Values are chi-square with one degree of freedom for changes in likelihood between successive steps in Miller and Spray's (1993) logistic discriminant function analysis. Identical values for two different variables indicate that response categories are the same for that item (see the Appendix).
* $p < .05$. ** $p < .01$. (Probability of obtaining data at least as extreme as those observed if there is no DIF. Strictly speaking, p values are higher because they are not corrected for multiple tests.)

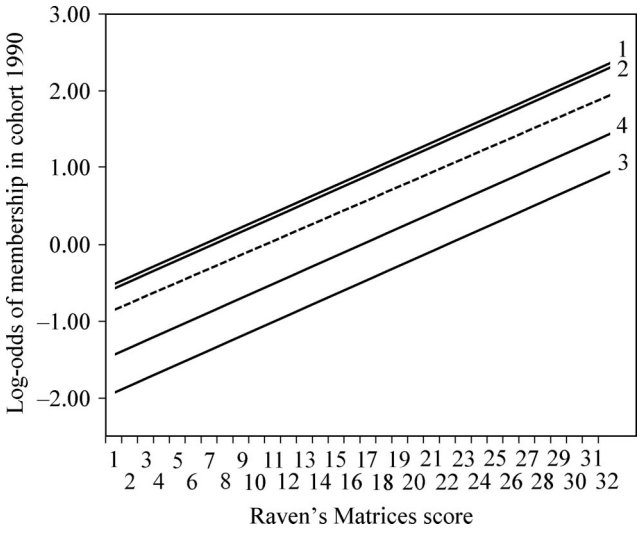

Figure 10. Logistic discriminant function analysis for differential item functioning (DIF) of Item 28 for the number-of-rules model. The *y*-axis represents the log-odds of Cohort 1990 membership when cohort membership is conditioned on Raven's Matrices score (the *x*-axis) and response category. Each solid line represents the function for one of the four response categories. The dotted line is the function of cohort membership conditioned solely on Raven's Matrices score (the "null" function). Thus, the solid lines overlap the dotted line when there is no DIF. The low position of higher response category functions in relation to the null function reveals that participants responding in the highest category at any given level of Raven's Matrices score are more likely to be members of Cohort 1940.

conditioned on Raven's Matrices score (the *x*-axis) and response category. Each solid line represents the function for one of the four response categories. The dotted line is the function of cohort membership conditioned solely on Raven's Matrices score (the "null" function). Thus, the solid lines overlap the dotted line when there is no DIF. The low position of higher response category functions in relation to the null function reveal that participants responding in the highest category at any given level of Raven's Matrices score are more likely to be members of Cohort 1940. Response patterns for this item are compatible with predictions in revealing that Cohort 1940 participants inferred more rules than Cohort 1990 participants who achieved the same score on the Raven's Matrices.

Eleven of the 18 affected items (Items 2, 3, 5, 6, 10, 14, 17, 19, 28, 31, and 36) have the same relatively straightforward interpretation as Item 28. The seven remaining items show the reverse effect, but six of these seven items have rules at Level 2 of dissimilarity or higher, making it impossible to rule out the confound between dissimilarity and number of rules as an alternative explanation, or at least a source of ambiguity that can only be resolved by determining which participants selected which responses from either category. In fact, response categories for two of these items (Items 7 and 34) are identical for the dissimilarity and number-of-rules variables, meaning that the incompatibility of their orders with our number-of-rules prediction constitutes support of our dissimilarity prediction. This is precisely why we made our predictions at the level of the test rather than individual items.

Indeed, an overall analysis of the entire set of 32 items, conditioning cohort membership on Raven's Matrices score, overall sum-score (the sufficient statistic for the latent variable rather than category for a single item) for the number-of-rules variable, and the interaction term, revealed an increased likelihood of Cohort 1940 membership for participants with high sum-scores relative to Raven's Matrices scores, $\chi^2(1) = 11.23$, $r = .23$. These results confirm our prediction that raw Raven's Matrices scores underestimate the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants.

Dissimilarity. An item-level analysis of response patterns verified the predicted direction of DIF for the dissimilarity variable in only four of eight items that revealed DIF (Items 7, 18, 30, and 34). Although an overall analysis across items revealed DIF in the direction that is opposite to predictions for the dissimilarity variable (sum-score), $\chi^2(1) = 4.37$, $r = .14$, this finding appears to be due entirely to overlap in response categories with the number-of-rules variable. The dissimilarity variable revealed no DIF when the number-of-rules variable was added as an additional criterion variable, $\chi^2(1) = 0.78$, $r = .06$. These results imply no overall DIF for the dissimilarity variable, but again, this conclusion is tentative because the variable is highly underdetermined by the items and response choices of the Raven's Matrices.

Graphical illustration. Some readers may find violations of measurement invariance more transparent in the familiar context of linear regression. Figure 11 is a linear regression-like scatterplot of raw Raven's Matrices scores as a function of scores on the latent variables. Polytomous sum-scores are displayed rather than thetas in keeping with a linear interpretation (thetas show the same basic effect in a logistic "S"-shape rather than a straight line). The figure shows that Raven's Matrices scores vary similarly between cohorts across levels of the dissimilarity variable, but tend to underestimate the number-of-rules variable for Cohort 1940 participants relative to Cohort 1990 participants as evidenced by the high proportion of white dots beneath the trend line. The size of the effect is highly constrained by the overlap in response categories between Raven's Matrices score and the number-of-rules variable, but the effect is nonetheless clearly visible. Consistent with predictions and our interpretation of Study 1, the Raven's Matrices test violates measurement invariance between cohorts by underestimating the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants. That is, for Cohort 1940 participants and Cohort 1990 participants who earn the same raw score, Cohort 1940 participants would correctly infer a greater number of rules.

Discussion

The purpose of Study 2 was to verify an assumption behind our interpretation of Study 1 while testing the prediction that Raven's Matrices scores overestimate the level of dissimilarity at which Cohort 1940 participants map objects relative to Cohort 1990 participants, or underestimate the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants.

The excellent fit of the data to the dichotomous Rasch model suggests that our interpretation of Study 1 pass rates as indicators of distribution-free difficulty is defensible, at least when considering young adults born sometime around or after 1940 in highly developed countries. Although constraints of test materials forced

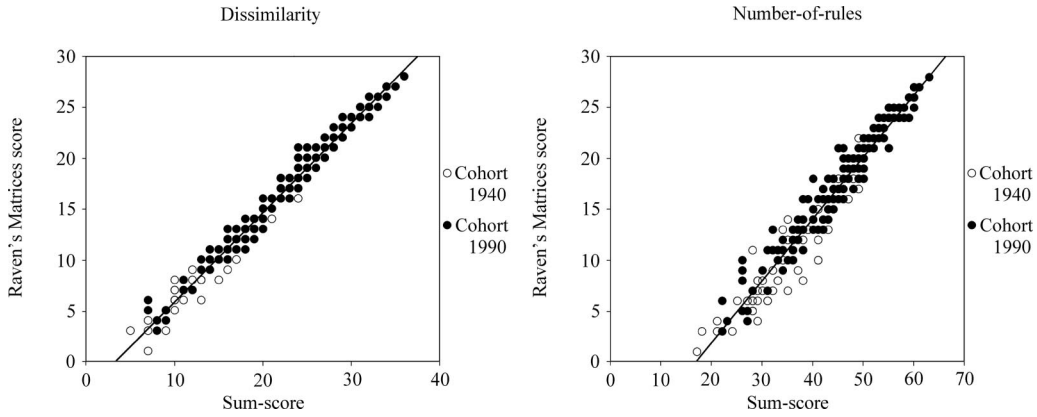

Figure 11. A linear regression-like depiction of raw Raven's Matrices scores as a function of placement along the latent variables. Sum-scores (sufficient statistics) are presented rather than thetas in keeping with a linear interpretation. The apparent scarcity of data is due to frequent overlap among the 223 cases. Raven's Matrices scores vary uniformly between cohorts across levels of the dissimilarity variable but tend to underestimate placement along the number-of-rules variable for Cohort 1940 participants relative to Cohort 1990 participants as evidenced by the frequency of white dots on the right side of the trend line.

both variables to correlate highly with raw test scores and with one another, results confirm our prediction that raw score underestimates the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants who achieve the same score on the Raven's Matrices.

It is important to note constraints that preexisting tests place on the degree to which scores on polytomous variables can deviate from raw score. Although response choices of Raven's Matrices items do vary in a manner that permits limited testing of predictions, the test includes very few items with more than two levels of dissimilarity in the response choices, which restricted the degree to which response categories for the dissimilarity variable could differ from mere accuracy. This lack of variation suppressed the opportunity to observe measurement violations of raw scores in relation to dissimilarity. This is probably why we did not find that Raven's Matrices scores overestimate the level at which Cohort 1940 participants map objects relative to Cohort 1990 participants even though this pattern of results would have been compatible with the findings of Study 1. The null finding does not rule out the possibility of observing the predicted effect for the level-of-dissimilarity variable with a test designed specifically to test the same predictions. Raven's Matrices items permitted the number-of-rules variable to deviate somewhat more from mere accuracy, revealing systematic DIF that is compatible with predictions despite methodological constraints of imperfect response categories.

For the same reason that the dissimilarity and number-of-rules variables overlap with raw score, they also overlap with one another. The study would have served little purpose if it were not possible in principle for participants to achieve different scores on the two variables, but the fact remains that it is impossible to achieve a high score on one variable without achieving a relatively high score on the other. For our purposes, it was necessary to use a test with a documented Flynn effect, ideally the same test completed by participants in Study 1, but given the constraints identified above, the most effective way of assessing measurement invariance in relation to theoretically-motivated variables in future studies is to design items whose response options vary systematically in accordance with predictions (e.g., Embretson, 1998; Freund et al., 2008).

It should not be forgotten that a difference between cohorts is, at least in this case, also a difference between age groups, but an age-related explanation of the findings is not easy to defend. As far as we know, there are no theories in cognitive aging that would make the same item-specific predictions as the current proposal. A cohort-related interpretation is more compatible with Study 1 findings, which confirmed that conceptually similar predictions with a data set of two cohorts that were comparable in mean age at the time of testing. Finally, and perhaps most decisively, the Flynn effect has to be caused by something that is distinct from causes of age-related cognitive decline. As we show below, the proposal that motivated predictions is highly compatible with the general pattern of between-cohort gains that is observed when various subtests are differentiated according to structure and content.

A final point merits special emphasis. The dichotomous model not only lends support to our distribution-free interpretation of Study 1 findings, but also illustrates an important limitation of interpreting latent variables as causal entities in their own right. The good fit of dichotomous data to the Rasch model along with the relative lack of DIF for dichotomous data could easily lead investigators to conclude that there exists an ability common to members of both cohorts . . . *if* it is forgotten that the Rasch model, like any other latent variable model, is a set of probabilistic criteria that does not arbitrate the existence or non-existence of psychological properties (Maraun, 1996). Because the dichotomous model defines ability as something no more specific than distribution-free patterns of response accuracy, it was never capable of distinguishing between any two theories of performance that are both compatible with distribution-free patterns of response accuracy.

In sum, our findings are compatible with the prediction that Raven's Matrices scores violate measurement invariance between cohorts by underestimating the number of rules inferred by Cohort 1940 participants relative to Cohort 1990 participants. A reason-

able interpretation of the findings is that raw score is, on average, constrained by different limiting factors in the two cohorts such that Cohort 1940 participants tend to lose more points than Cohort 1990 participants because of their inability to map dissimilar objects. A more general (and perhaps less cautious) interpretation is that test-takers born around 1940 are more limited than their recent-born counterparts in their ability to form abstract concepts, but not in their ability to keep track of these concepts once they are formed. Regardless of interpretation, our results imply that there is nothing paradoxical about Cohort 1940 participants in Study 1 having achieved the same overall pass rates as Cohort 1990 participants despite having lower pass rates on items with dissimilar corresponding objects.

Confirmatory factor-analysis has already shown that test batteries violate measurement invariance between cohorts (Must et al., 2009; Wicherts et al., 2004). The present study is the first to confirm a prediction of measurement invariance within a single preexisting test using latent variables defined by a cognitive account of rising scores.

General Discussion

Studies 1 and 2 provide converging support for our proposal that rising scores reflect improved mapping of dissimilar analogical objects. The findings are of immediate relevance to the cognition of matrix reasoning and have important implications for cognitive aging research. We conclude by placing the findings in the larger context of rising scores.

Implications for Matrix Reasoning and Cognitive Aging

Unlike others who have investigated matrix reasoning, we do not attribute our findings to individual differences in working memory, but instead note that our proposal is compatible with Carpenter et al.'s (1990) well-known findings precisely because their FAIRAVEN and BETTERAVEN models *differ only in productions* (procedural knowledge). More specifically, BETTERAVEN's additional productions enable the model to recognize rules containing dissimilar objects. In ordinary language, 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*). In contrast, BETTERAVEN's additional productions allow it to test other rules when the mapping of matching names does not successfully elicit a rule. Although neither model *infers* rules per se, BETTERAVEN's advantage over FAIRAVEN is highly compatible with the thesis of this article.

Attributing differences in test scores to differences in working memory requires a definition of working memory that is logically distinct from performance itself (e.g., see Boag, 2011; Maraun, 1998; Maze, 1954; Michell, 2011; Wallach & Wallach, 1998). Carpenter et al. (1990) provided such a definition by formalizing working memory demand in terms of the features of items. However, the working memory demand of items, as defined by Carpenter et al.'s models, does not predict the magnitude of correlations between working memory span scores and item-level accuracy on the Raven's Matrices (Unsworth & Engle, 2005; see also Wiley, Jarosz, Cushen, & Colflesh, 2011). That an accepted theoretical construct of working memory demand (a prospective source of item difficulty) is incompatible with an accepted empirical construct of working memory (observed ability as defined by performance on a working memory test) is testimony to the indeterminacy of the term, "working memory," as it is currently used in the literature. Our conclusions can neither corroborate nor contradict working memory claims until investigators agree on a definition of working memory that enables claims about the construct to be disconfirmed, that is, a single, a priori criterion for employment of the term "working memory."

Studies 1 and 2 are consistent with earlier conclusions that number of rules is a source of item difficulty (Carpenter et al., 1990; Embretson, 1998). However, preserving rules requires mapping objects in the first place, which is why accounting for individual differences entails not just identifying variables, but ascertaining how various levels of performance are achieved (Borsboom, Mellenbergh, & van Heerden, 2004; Ericsson & Kintsch, 1995). Study 2 is an empirical demonstration of why it is problematic to equate observed differences in scores, including latent variable scores, with literal psychological quantities.

For very similar reasons, our findings encourage cognitive aging researchers to be cognizant of the Flynn effect and its implications. The trend constitutes a major cross-sectional confound that is seldom mentioned in this literature. Dickinson and Hiscock (2010) concluded after analyzing normative data from two versions of the WAIS—Wechsler Adult Intelligence Scale—Revised (WAIS–R; Wechsler, 1981) and Wechsler Adult Intelligence Scale—Third Edition (WAIS–III; Wechsler, 1997)—that cohort is responsible for the *majority* of the differences in cross-sectional scores obtained across subtests for groups separated by 50 years of age. In an earlier study, Hiscock (2007) estimated that only about one third of the cross-sectional difference in Raven's Matrices scores is attributable to age. The present findings lend substance to concerns raised by others (Hofer & Sliwinski, 2001; Schaie, 2009; Zelinski & Kennison, 2007) that effects of cohort and time period are understated or misrepresented by prevailing interpretations of cross-sectional findings.

Making Sense of the Flynn Effect

The gold standard for any theory of rising scores is accounting for gains on Raven's Matrices. However, our emphasis on this test has led us to understate the application of our proposal to other tests that are seemingly less abstract. It is informative to return to Flynn and Weiss's (2007) discussion of Similarities.

Assuming children are familiar with *dusk* and *dawn*, presentation of these two concepts would tend to activate other concepts common to both. *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, a child who knows to expand her search beyond the obvious can evaluate further possibilities. If she retrieves both *time of day* and *intermediate brightness*, she can treat them as objects in need of roles and perhaps, infer 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 full-fledged role (*time of day*) as an object in need of a more abstract role (*separates night and day*). This does not imply that she would not benefit from

additional knowledge (e.g., a heuristic of attempting to account for the most objects with the fewest relations). 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 greater representational flexibility must disappear in the presence of content.

If the ability to map objects between items has contributed to higher scores, gains should be largest on tests composed of items with a structure that is both initially unfamiliar and relatively uniform from item to item. Knowing how to cope with indeterminacy 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 Wechsler and Stanford–Binet both contain many subtests requiring problem solving procedures that test-takers would seldom encounter outside the context of intelligence testing, and that remain relatively consistent throughout an individual test. Consistent with predictions, these tests show moderate improvement across subtests.

The lowest gains are observed on subtests consisting of items that resemble 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 their structures were unfamiliar, the items call for 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 are unfamiliar to most test-takers.

In general, the theory predicts that gains in raw scores should be highest on tests where higher-level analogical mapping 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 respond correctly. This question can only be answered by task analysis (e.g., Ericsson & Simon, 1993).

Cross-Cultural Implications

Our proposal that improved test performance reflects a form of knowledge that proliferates only in modern cultures is consistent with Brouwers et al.'s (2009) cross-cultural meta-analysis of the Raven's Matrices. This analysis revealed that scores at any given time (i.e., when controlling for publication year) were 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.

Conceived in very simple terms, possession of a form of knowledge will correlate with performance on various tests and other tasks to the extent that it facilitates performance *and* is neither too common nor scarce within a population (see Wicherts & Johnson, 2009). Thus, psychometric properties of items and tests, such as their covariation with other tests (i.e., their so-called *g*-loadings), will be lowest when either very few or very many people have acquired the knowledge, and highest when about half the population has acquired it. By this reasoning, our proposal is compatible with Wicherts et al.'s (2010) exhaustive analysis of Raven's Matrices scores of sub-Saharan Africans, which revealed relatively low *g*-loadings in this population of test-takers who are unlikely to have acquired a form of knowledge that is conferred only by modern cultures. By the same reasoning, our proposal is compatible with a decline in covariance over time (Kane & Oakland, 2000) in the United States where the knowledge has become a standard feature of higher-level cognition.

Findings of Studies 1 and 2 are compatible with a growing literature revealing violations of measurement invariance between cohorts (Beaujean, & Osterlind, 2008; Must et al., 2009; Wicherts et al., 2004). However, Study 2 is also a demonstration of why accuracy data cannot be expected to reveal violations of measurement invariance in terms of how responses are generated when two distinct approaches to generating responses both confer distribution-free patterns of accuracy. In other words, two populations may not be comparable to one another even when measurement invariance is observed if the research question one seeks to answer by comparing these populations is more specific than the data used to establish invariance. Ultimately, one cannot rule out violations of measurement invariance entirely, but only attempt to test increasingly detailed hypotheses about how various levels of performance are achieved in two or more populations.

Summary

This article attempts to account for rising scores on culture-free intelligence tests as a knowledge-based phenomenon by reconciling Flynn's (2007) proposal that rising scores were caused by improved abstract reasoning with insights and discoveries that have emerged from studies of matrix reasoning (e.g., Carpenter et al., 1990; Embretson, 1998; Primi, 2002; Meo et al., 2007).

A review of the literature suggests that the level of dissimilarity at which individuals map 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 predecessors of 50 years ago. Polytomous Rasch models suggest that Raven's Matrices scores of today's young adults are constrained less by the inability to map dissimilar objects than scores of young adults from around 1960.

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.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences, 36,* 1459–1470. doi:10.1016/S0191-8869(03)00241-1

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23. doi:10.1177/0146621697211001

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42,* I-7–I-16. doi:10.1097/01.mlr.00000103528.48582.7c

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54,* 394–403. doi:10.1177/0013164494054002013

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

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence, 36,* 455–463. doi:10.1016/j.intell.2007.10.004

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

Boot, W. R., Champion, M., Blakely, D. P., Wright, T., Souders, D. J., & Charness, N. (2012). *Video game interventions as a means to address cognitive aging: Perceptions, attitudes, and effectiveness*. Manuscript submitted for publication.

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

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

Bunge, M. (1997). Mechanism and explanation. *Philosophy of the Social Sciences, 27,* 410–465. doi:10.1177/004839319702700402

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

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence, 4,* 185–211. doi:10.1080/09528139208953747

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

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

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

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., & 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. (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*. doi:10.1017/CBO9780511605253

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. doi:10.1111/j.2044-8279.1964.tb00632.x

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

Freund, P. A., Hofer, S., & Holling, H. (2008). Figural matrix items explaining and controlling for the psychometric properties of computer-generated matrix items. *Applied Psychological Measurement, 32,* 195–210. doi:10.1177/0146621607306972

Gallini, J. K. (1983). A Rasch analysis of Raven item data. *Journal of Experimental Education, 52,* 27–32.

Green, K. E., & Kluever, R. C. (1992). Components of item difficulty of Raven's Matrices. *Journal of General Psychology, 119,* 189–199. doi:10.1080/00221309.1992.9921172

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

Hidalgo, M. D., & Gomez, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial regression: A comparison for polytomous items. *Quality & Quantity, 40,* 805–823. doi:10.1007/s11135-005-3964-2

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

Kane, H., & Oakland, T. D. (2000). Secular declines in Spearman's g: Some evidence from the United States. *The Journal of Genetic Psychology: Research and Theory on Human Development, 161,* 337–345. doi:10.1080/00221320009596716

Kelderman, H. (1996). Multidimensional models for partial-credit scoring. *Applied Psychological Measurement, 20,* 155–168. doi:10.1177/014662169602000205

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

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

Linhares, A. (2000). A glimpse at the metaphysics of Bongard problems. *Artificial Intelligence, 121,* 251–270. doi:10.1016/S0004-3702(00)00042-4

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, August 27). A long-term increase in the fluid intelligence of English children. *Nature, 328,* 797. doi:10.1038/328797a0

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

Maraun, M. D. (1996). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research, 31,* 517–538. doi:10.1207/s15327906mbr3104_6

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8,* 435–461. doi:10.1177/0959354398084001

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. doi:10.1007/BF02296272

Maze, J. R. (1954). Do intervening variables intervene? *Psychological Review, 61,* 226–234. doi:10.1037/h0061026

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

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function-analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107–122. doi:10.1111/j.1745-3984.1993.tb01069.x

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 461–473. doi:10.1007/s11336-007-9039-7

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 Nijenhuis, 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

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

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

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12,* 220–229. doi:10.1111/j.0965-075X.2004.00276.x

Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology, 84,* 171–199. doi:10.1111/j.2044-8295.1993.tb02472.x

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

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. doi:10.1016/0191-8869(91)90287-L

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). doi:10.1037/10270-005

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8,* 33. doi:10.1186/1471-2288-8-33

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18,* 313–350. doi:10.1207/s15324818ame1804_1

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

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.2005.01.029

te Nijenhuis, J., Murphy, R., & van Eeden, R. (2011). The Flynn effect in South Africa. *Intelligence, 39,* 456–467. doi:10.1016/j.intell.2011.08.003

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 der Ven, A., & Ellis, J. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29,* 45–64. http://dx.doi.org/10.1016/S0191-8869%2899%2900177-4

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

Wallach, M. A., & Wallach, L. (1998). When experiments serve little purpose: Misguided research in mainstream psychology. *Theory & Psychology, 8,* 183–194. doi:10.1177/0959354398082005

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York, NY: Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *The Wechsler Adult Intelligence Scale—Third Edition*. San Antonio, TX: Psychological Corporation.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40,* 73–76. doi:10.1016/j.intell.2012.01.004

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn effect. *Learning and Individual Differences, 20,* 135–151. doi:10.1016/j.lindif.2009.12.001

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

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings of the Royal Society B: Biological Sciences, 276,* 2675–2683. doi:10.1098/rspb.2009.0238

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven's Advanced Progressive Matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 256–263. doi:10.1037/a0021613

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: MESA.

Yates, A. J. (1961). Item analysis of progressive matrices: 1947. *British Journal of Educational Psychology, 31,* 152–157. doi:10.1111/j.2044-8279.1961.tb02927.x

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

(*Appendix follows*)

Appendix A

Item Classifications for Study 1 and Polytomous Response Categories for Study 2

	Study 1		Study 2			
			Dissimilarity		Number-of-rules	
Item	Dissimilarity	No. of rules	Category	Responses	Category	Responses
1	1.67	3	1	(2, 3, 6, 7, 8)	1	(6, 7)
			2	(1, 4)	2	(1, 2, 3, 4, 8)
			3	(5)	3	(5)
2[a]	1.00	2	1	(2, 3, 4, 5, 6, 7, 8)	1	(4, 7)
			2	(1)	2	(2, 3, 5, 6, 8)
					3	(1)
3	1.00	2	1	(1, 2, 3, 4, 5, 6[d], 7, 8)	1	(1, 4, 5, 6[d])
			2	(7)	2	(2, 3, 8)
					3	(7)
4	1.00	2	1	(1, 2, 3, 5, 6, 7, 8)	1	(2, 6, 7, 8)
			2	(4)	2	(1, 3, 5)
					3	(4)
5	1.00	2	1	(1, 2, 4, 5, 6, 7, 8)	1	(6)
			2	(3)	2	(1, 2, 4, 5, 7, 8)
					3	(3)
6	1.00	2	1	(2, 3, 4, 5, 6, 7, 8)	1	(4, 5, 6, 7)
			2	(1)	2	(2, 3, 8)
					3	(1)
7	2.00	1	1	(1, 2, 3, 4, 5, 7, 8)	1	(1, 2, 3, 4, 5, 7, 8)
			2	(6)	2	(6)
8	2.00	2	1	(2, 3, 4, 6, 7, 8)	1	(2, 5, 6, 7, 8)
			2	(1)	2	(3, 4)
					3	(1)
9	2.00	2	1	(1, 2, 3, 4, 5, 6, 7)	1	(1, 2, 3, 4, 5, 6, 7)
			2	(8)	2	(8)
10[c]	1.00	2	1	(1, 2, 3[d], 5, 6, 7[d], 8)	1	(3[d], 5, 6, 7[d])
			2	(4)	2	(1, 2, 8)
					3	(4)
11[a]	2.00	1	1	(1, 2, 3, 4, 6, 7, 8)	1	(1, 2, 3, 4, 6, 7, 8)
			2	(5)	2	(5)
12	2.00	1	1	(1, 2, 3, 4, 5, 7, 8)	1	(1, 2, 3, 4, 5, 7, 8)
			2	(6)	2	(6)
13	1.33	3	1	(4)	1	(4)
			2	(1, 3, 5, 6, 7, 8)	2	(1, 8)
			3	(2)	3	(3, 5, 6, 7)
					4	(2)
14[c]	1.00	2	1	(2, 3, 4, 5, 6[d], 7, 8)	1	(2, 3, 4, 5, 6[d])
			2	(1)	2	(7, 8)
					3	(1)
15[a]	2.00	2	1	(1, 3, 4, 5, 6, 7, 8)	1	(1)
			2	(2)	2	(3, 4, 5, 6, 7, 8)
					3	(2)
16	2.00	1	1	(1, 2, 3, 5, 6, 7, 8)	1	(1, 2, 3, 5, 6, 7, 8)
			2	(4)	2	(4)
17	1.50	2	1	(1, 2, 4, 5, 7)	1	(1, 2, 4, 5, 7)
			2	(3, 8)	2	(3, 8)
			3	(6)	3	(6)
18[ab]	2.00	2	1	(1, 2, 3, 4, 5, 6, 8)	1	(2, 4, 5, 6)
			2	(7)	2	(1, 3, 8)
					3	(7)
19[ab]	2.00	2	1	(1, 2, 4, 5, 6, 7, 8)	1	(1, 2, 6, 8)
			2	(3)	2	(4, 5, 7)
					3	(3)

(*Appendix continues*)

Appendix (*continued*)

	Study 1		Study 2			
			Dissimilarity		Number-of-rules	
Item	Dissimilarity	No. of rules	Category	Responses	Category	Responses
20[a]	2.00	1	1	(1, 2, 3, 4, 5, 6, 7)	1	(1, 2, 3, 4, 5, 6, 7)
			2	(8)	2	(8)
21[ac]	2.00	4	1	(1, 2, 3, 4, 5[d], 6[d], 7[d])	1	(4, 5[d], 6[d], 7[d])
			2	(8)	2	(2, 3)
					3	(1)
					4	(8)
22	3.00	3	1	(1, 2, 3, 4, 5, 6, 8)	1	(1, 4, 6)
			2	(7)	2	(8)
					3	(2, 3, 5)
					4	(7)
23	3.00	4	1	(1, 2, 3, 4, 5)	1	(4, 7)
			2	(6)	2	(1)
					3	(2, 3, 5, 8)
					4	(6)
24[a]	1.00	2	1	(1, 2[d], 4, 5, 6, 7, 8)	1	(2[d], 7)
			2	(3)	2	(1, 4, 5, 6, 8)
					3	(3)
25[ac]	2.00	3	1	(1, 2, 3, 4, 6, 8)	1	(1, 2, 3, 4, 5, 6, 8)
			2	(5)	2	(7)
			3	(7)		
26	3.00	2	1	(1[d], 3[d], 4[d], 5, 6, 7, 8[d])	1	(1[d], 3[d], 4[d], 7, 8[d])
			2	(2)	2	(5, 6)
					3	(2)
27	3.00	2	1	(1, 2, 3[d], 4, 5[d], 6, 8[d])	1	(2, 3[d], 5[d], 8[d])
			2	(7)	2	(1, 4, 6)
					3	(7)
28[a]	2.00	4	1	(1, 2, 3, 4, 6, 7, 8)	1	(1)
			2	(5)	2	(3, 6, 7)
					3	(2, 4, 8)
					4	(5)
29[c]	2.33	3	1	(1)	1	(1)
			2	(2, 3, 4, 5, 6, 7, 8)	2	(3, 4, 5)
			3	(6)	3	(2, 7, 8)
					4	(6)
30[a]	3.00	3	1	(1, 2, 3, 4, 6, 7, 8)	1	(1, 2, 7)
			2	(5)	2	(3, 6, 8)
					3	(4)
					4	(5)
31	2.67	4	1	(1, 3, 7, 8)	1	(5, 7)
			2	(2, 5, 6)	2	(1, 2, 3, 6, 8)
			3	(4)	3	(1)
					4	(4)
32	2.33	4	1	(2, 6)	1	(6)
			2	(1, 3, 4, 5, 7)	2	(1, 2, 3)
			3	(8)	3	(5, 7)
					4	(4)
					5	(8)
33[c]	2.00	2	1	(1, 2, 3, 4, 6, 7, 8)	1	(1, 3, 4, 6, 7, 8)
			2	(5)	2	(2)
					3	(5)
34	2.25	4	1	(2, 4, 6, 7, 8)	1	(2, 4, 6, 7, 8)
			2	(3, 5)	2	(3, 5)
			3	(1)	3	(1)

(*Appendix continues*)

Appendix (*continued*)

	Study 1		Study 2			
			Dissimilarity		Number-of-rules	
Item	Dissimilarity	No. of rules	Category	Responses	Category	Responses
35	2.75	4	1	(1, 2[d], 4[d], 5[d], 6[d], 8)	1	(1, 2[d], 4[d], 5[d], 6[d], 8)
			2	(7)	2	(7)
			3	(3)	3	(3)
36	2.80	5	1	(1, 4, 6, 7)	1	(1, 6)
			2	(3, 5, 8)	2	(4, 5)
			3	(2)	3	(3, 7)
					4	(2)

Note. Category = ordinal rank of response with respect to latent variable. The number of categories correspond to Carpenter et al.'s (1990) study and Studies 1 and 2 to the extent permitted by response choices.
[a] Carpenter et al. (1990) did not report their own classification of item. [b] Item cannot be classified based on Carpenter et al.'s (1990) taxonomy (see p. 431 of their article). [c] Item was not analyzed in Study 2 because responses were not available for every participant (see the Method section). [d] Response has been placed in lowest category because it includes an incorrect object that is incompatible with rules as defined in the article.

Received May 28, 2012
Revision received August 14, 2012
Accepted August 14, 2012 ■