Behavioral Development Bulletin
2016, Vol. 21, No. 1, 50–62

© 2016 American Psychological Association
1942-0722/16/$12.00 http://dx.doi.org/10.1037/bdb0000015

An Analysis of the Verbal Comprehension Index of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV) Using the Model of Hierarchical Complexity (MHC): Why Might Stage Be a Better Measure of "Smarts" Than Verbal IQ?

Kyle Gramer Featherston
Dare Institute, Cambridge, Massachusetts

Shuling Julie Chen
New York University

Maria Toth-Gauthier and James M. Day
Université catholique de Louvain

Philippe Herman, Raquel Laverdeur, Laurence Nicolaï, Anne-Catherine Nicolay, Alice Sini, Caroline Tilkin, Mireille Tyberghein, and Christian Vanheck
Le Centre pour la Valorisation des Intelligences Multiples, Liège, Belgium

While IQ tests are the most common and largely accepted measurement of how "smart" a person is, whether they are the best measure of this construct is up for debate. This paper will discuss the relationship between IQ tests and their corresponding order of hierarchical complexity developmental stage scores based on the model of hierarchical complexity (MHC). The Verbal Comprehension Index (VCI) scales of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV) were used for scoring. The study shows that, according to the Hierarchical Complexity Scoring System (HCSS), the WAIS–IV fails to test verbal intelligence beyond the formal stage. This study used Rasch analysis to demonstrate that scoring the VCI of the WAIS–IV from a developmental sequence using the HCSS was successful in explaining the majority of the difficulty in VCI items. Much of the additional difficulty of tasks came from the knowledge of rare items and noise. This demonstrates the ceiling effect of the VCI of the WAIS–IV. Difficulties with scoring items, additional limitations with the IQ test, and their implications are discussed.

Keywords: intelligence, IQ, model of hierarchical complexity, development, testing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

For years, psychologists have attempted to define and test the concept of intelligence. The field has created and revised theories of intelligence, and countless instruments have been designed to attempt to match these theories. Still today there exist many theories of intelligence. The most influential of these posit a general cognitive ability or *g factor*. While this g factor, according to the Cattell–Horn–Carroll theory of multiple intelligences, is made up of several factors—most significantly, fluid and crystallized intelligence—the idea of the existence of an overlying general cognitive ability is very prominent in the field today. IQ is supposed to

Kyle Gramer Featherston, Dare Institute, Cambridge, Massachusetts; Shuling Julie Chen, Department of Psychology, New York University; Maria Toth-Gauthier and James M. Day, Department of Psychology and Educational Sciences, Université catholique de Louvain; Philippe Herman, Raquel Laverdeur, Laurence Nicolaï, Anne-Catherine Nicolay, Alice Sini, Caroline Tilkin, Mireille Tyberghein, and Christian Vanheck, Le Centre pour la Valorisation des Intelligences Multiples, Liège, Belgium.

Michael Lamport Commons discussed this topic with the authors and provided notes on which this paper is partially based.

Correspondence concerning this article should be addressed to Kyle Gramer Featherston, 234 Huron Avenue, Cambridge, MA 02138. E-mail: kylegramerfeatherston@gmail.com

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

be a rough estimate of this general ability (Kamphaus, Winsor, Rowe, & Kim, 2005). IQ is a largely accepted construct in today's psychometric research community. In fact, IQ is often used interchangeably with the terms *general cognitive ability*, *mental ability*, and *intelligence*, both in academics to describe the common core that cognitive tests share (Deary, Penke, & Johnson, 2010) and in everyday common conversation.

The modern IQ test evolved from an intelligence test that French psychologist Alfred Binet and colleague Theodore Simon developed in order to identify students with learning disabilities (Binet & Simon, 1914). Binet viewed the test as a measure of scholastic ability and did not believe that it was a measure of intelligence, nor that intelligence was a singular construct that could be identified in such an instrument. He thus condemned those for using it as such a measure (White, 2000). There are several different IQ tests that are used based on variations of intelligence models, but the most common instrument used with adults is the Wechsler Adult Intelligence Scale (WAIS). The most recent version is the fourth edition (WAIS–IV). However, IQ tests have received criticisms from different areas of psychology for many years (Borsboom, 2006; Mackintosh, 2011; McClelland, 1973; Neisser et al., 1996; Schönemann, 1997), and their predictive ability of success has been called into question. Whether or not IQ tests are an accurate measure of "smartness" has been a hotly debated topic in psychology for a long time. This paper will argue that the idea of IQ in general, specifically through the WAIS–IV, does not measure complex problem solving of multiple variables, as defined by a quantitative behavioral developmental theory, the model of hierarchical complexity (MHC).

There are a great number of flaws with the WAIS–IV and similar tests of IQ, although these tests are used broadly by many psychologists for various purposes. IQ tests are developed using norms and *psychometric analysis*. The responses are only analyzed in psychometrics using factor analysis. The problem with this is that the test is not developed with an idea of what is being measured in mind or with an a priori knowledge of what items will be more difficult than others and why. There is no stimulus measure in this type of analysis. The analysis only identifies which questions are difficult after the fact. The analysis misses the characteristic of the stimuli—in this case, the test items—that cause the discrepancy in responses. To a large extent, intelligence tests only test surface information (McClelland, 1973). Additionally, these tests have been shown to have cultural and education-level biases. The information subtest, for example, depends on culture, experience, and knowledge to a large extent. People in other cultures may not have knowledge of specific facts the WAIS–IV tests, which may lower their scores simply based on their past experiences. While the creators argue that these are tests of "general cognitive ability," what is actually tested in many instances is whether or not certain knowledge has been learned in a participant's educational or cultural history. McClelland (1973) argues that intelligence tests may mainly predict test taking and symbol manipulation competencies as opposed to actual "smartness." Recent additions to the test have included measurement of basic cognitive processes such as working memory and processing speed. Instead, the everyday notion of "smartness" may be better defined by the complexity of a task a person is able to accomplish.

An alternative way to look at intelligence is to view it as a progression along a developmental sequence. Inhelder and Piaget's (1958) definition of IQ is an individual's place in a universal sequence of development toward formal operational reasoning. The Piagetian cognitive stage measures provide a rational standard for educational intervention (Kohlberg & Mayer, 1972). Piaget's definition of intelligence is not limited to school-type success (Devries, 1974) but instead takes the long-range perspective of the evolution of knowledge and intelligence in the individual. It describes changes with age in the structure of knowledge and changes in reasoning about reality. There have been several studies that have attempted to determine the correlation between traditional IQ measures and Piagetian developmental tasks. These studies have had a wide range of results, with the average correlation being a modest $r = .578$ (see Table 1).

While these were compelling studies, their flaw is that Inhelder and Piaget's (1958) developmental task sequence is limited. While their developmental model ended at formal operations, future research has demonstrated that

Table 1
Correlations Between Piagetian Tasks and IQ Tests

Correlation	Tests used	*N*	Age
r = .34[a]	Fifteen Piaget-type tasks and Stanford–Binet IQ Test	143	5–7
r = .837[b]	Twenty-seven heterogeneous Piagetian tasks and 11 Wechsler subtests	150	6–10, 10–14, and 14–18
r(kindergarten)[c] = .48 r(Grade 1) = .52 r(Grade 2) = .56	Nine Piaget tests and Wechsler Intelligence Scale for Children	100	5–8

[a] Adapted from "Relationships Among Piagetian, IQ, and Achievement Assessments," by R. DeVries, 1974, *Child Development, 45,* pp. 746–756. Copyright 1974 by Wiley. [b] Adapted from "Piagetian Tasks Measure Intelligence and Intelligence Tests Assess Cognitive Development: A Reanalysis," by L. G. Humphreys and C. K. Parsons, 1979, *Intelligence, 3,* pp. 369–381. Copyright 1979 by Elsevier. [c] Adapted from "Relationship of Piaget Measures to Standard Intelligence and Motor Scales," by S. Z. Dudek, E. P. Lester, J. S. Goldberg, and G. B. Dyer, 1969, *Perceptual and Motor Skills, 28,* pp. 351–362. Copyright 1969 by Sage.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

adults are capable of reasoning beyond this level (Commons, Richards, & Kuhn, 1982). It is through this reasoning that the MHC was developed.

The MHC

The MHC is a nonmentalistic, neo-Piagetian, and quantitative behavioral development theory. It offers a standard method of examining the universal pattern of development. A fundamental assumption is that development proceeds across a large number of general sequences of behavior. These sequences exist in every domain, including, but not limited to, the mathematical, logical, scientific, moral, social, and interpersonal domains. The stages of the MHC have been shown to predict humans' "smartness" in the colloquial sense using the laundry and balance beam instruments (Commons et al., 2008).

The different layers in a hierarchical sequence of task complexity are referred to as *orders*. The successful completion of a task of a given order is referred to as a *stage*. Orders of hierarchical complexity (OHCs) assess the predicted difficulty of behavior tasks (Commons, Gane-McCalla, Barker, & Li, 2014; Commons & Miller, 1998; Commons & Pekker, 2008; Commons & Richards, 1984; Commons, Trudeau, Stein, Richards, & Krause, 1998). The OHC is an equally spaced unidimensional ordinal scale that measures difficulty independent of domain and content. The higher the OHC, the more difficult the task. The order names and numbers are listed in Table 2.

There are three axioms of the MHC. A higher order action (a) is defined in terms of task actions from the *next lower* order of hierarchical complexity, (b) *organizes* two or more less complex actions, and (c) is carried out in a *nonarbitrary* way (see Figure 1).

Why Use the MHC to Score IQ Tests?

As discussed previously, there are a number of flaws with traditional IQ tests, including the WAIS–IV. While no proposed view of intellectual ability is without flaws, the MHC has a few advantages over traditional IQ testing. First, IQ tests generally rely largely on testing previously

Table 2
Orders of Hierarchical Complexity

Order number	Order name
0	Computational
1	Automatic
2	Sensory or motor
3	Circular sensory motor
4	Sensory motor
5	Nominal
6	Sentential
7	Preoperational
8	Primary
9	Concrete
10	Abstract
11	Formal
12	Systematic
13	Metasystematic
14	Paradigmatic
15	Crossparadigmatic
16	Meta-crossparadigmatic

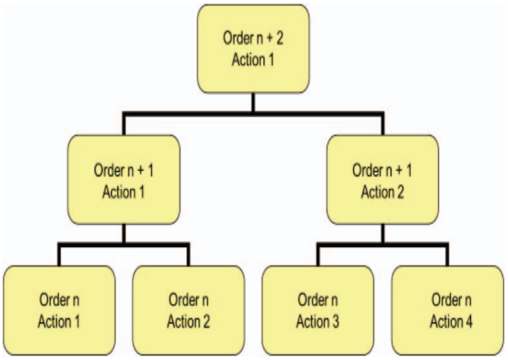

Figure 1. Task complexity. Each higher order is demonstrated by a combination of two (or more) tasks from the next lowest order in a nonarbitrary way. This figure demonstrates how an order $n + 2$ action is defined by the two actions from order $n + 1$, which are themselves defined by a nonarbitrary combination of two order n actions.

learned knowledge. On the other hand, the MHC is cultural, content, and education free. Second, the MHC does not rely on psychometrically analyzed norms but is instead based on a simple, clear mathematical model. While using psychometrics is not a problem, there is a lack of knowledge of the mechanism behind the difficulty of items in IQ tests. Using the MHC, it is possible to determine the OHC of a task a priori. The MHC uses a psychophysical approach in conjunction with psychometrics. Psychophysics, one of the branches of behavioral science, is the study of quantitative relations between psychological events and physical events or, more specifically, between sensations and the stimuli that produced the sensations (Pelli & Farell, 1995). The initial step toward defining psychophysically answerable questions is to formulate the problem as a task that the observer must perform. The psychophysical approach is to find the properties of the items of a test to predict the difficulty of performance. Using this approach, it is possible to have a better understanding of the complexity of tasks a person is either completing successfully or failing to complete instead of giving a raw score that puts a person on a spectrum of "general ability" with no insight into what that means.

This study will be unique in its use of psychophysical principles in assessing IQ testing. Based on this approach, items from the WAIS–IV Verbal Comprehension Index (VCI) will be scored based on their OHC. Based on these scores, it should be possible to predict the items that are of a higher difficulty. Additionally, by seeing which participants complete tasks at each order, the stages of the participants will be estimated. The estimated stages will then be compared to IQ scores to see if and how the two measures of smarts are related.

Method

Participants

The participants were French-speaking participants ($N = 101$, 54 female, 47 male) from Le Centre pour la Valorisation des Intelligences Multiples (the Center for the Valuation of Multiple Intelligences) in Liège, Belgium. The Center works to support "gifted" people; thus, a high-IQ sample was expected. The participants were all considered adults ranging from 16 to 68 years old ($M = 30.6$, $SD = 11.934$).

Instruments

Participants were given the French edition of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV). The WAIS–IV includes four indices and a Full Scale IQ (FSIQ) score (Wechsler, 2008). The indices are the VCI, the Perceptual Reasoning Index, the Working Memory Index, and the Processing Speed Index. The test is based on theories of cognitive abilities, with the four indices posited to make up some of the major factors that influence g, with full-scale IQ approximating the g factor. In total, there are 10 core subtests and five supplemental subtests that make up the test, with each subtest loading on one of the indices. For most of the analysis in this study, the three subtests that make up the VCI—the Vocabulary (Vocabulaire), Similarities (Similitudes), and Information (Information) subtests—were used. The analysis was run using SPSS Predictive Analytics Software (PASW) Statistics 18 and Winsteps Version 3.74.0.

Procedure

The WAIS–IV was given to the participants as part of a normal evaluation by the counselors at the center using the guidelines laid out in the scoring guide. The only difference was that for the vocabulary similarities and information sections, the evaluator recorded the points received

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

for each item. The vocabulary section asks participants to define words, and based on how accurate their definitions are according to the scoring guide, they are given 0, 1, or 2 points. The similarities section asks participants to compare how two words are similar, and they are also given 0, 1, or 2 points based on how accurate their answers are according to the scoring guide.

Items were translated directly from French to English. In order to preserve the meaning of the items, only those items that were identical in French and English were used for analysis. The items were scored independently by three groups of scorers using the Hierarchical Complexity Scoring System (HCSS; Commons, Miller, Goodheart, & Danaher-Gilpin, 2005). The HCSS is a system to score the difficulty of tasks based on the order of hierarchical complexity. It entails several steps for assessing performance on a task. Any differences in scoring were resolved in a group discussion among all raters. The raters were able to come to a consensus on all items. Twenty items were scored for both their 2- and 1-point responses. Zero-point responses were considered random and were not scored.

A second measure was taken to evaluate the difficulty of items, which was evaluating the rarity of the words used in the verbal section. Each word was reviewed based on the frequency of appearance in Belgian works in the French language. Each word that was included as one of the 10,000 most common words in the French language was given a score of 0, and those that were not were given a score of 1.

Results

The mean IQ of the sample was 121.34 (*SD* = 12.497), well above the population mean of 100. On the items tested, the participants scored a range of 61 to a perfect score of 80 out of 80, with an extremely high mean of 72. 87(*SD* = 4.237). Four participants got the full 2 points on every question we tested, and several more got close-to-perfect scores, with six getting 1 point short of a perfect score. There were also five questions that every participant received 2 points on and an additional three that every participant got at least 1 point on.

The translation of the items yielded 20 items from the Vocabulary and Similarities subtests that were considered exact translations from French to English. Only using items that were exactly the same in English and French avoided any language confusion in the scoring manual. The information subsection lacked items that directly translated to the English version and also asked for simple recall of information. This was ruled to be entirely concrete stage tasks with no higher stage. For these reasons, items from the information section were not used in the analysis. For the purpose of the analysis, 1-point responses were considered separate questions from 2-point responses. A correct 2-point response was considered a correct answer on the 1-point item. This left 40 items. Table 3 shows the sample scoring of an item from the similarities section, with three raters all in agreement.

As stated previously, a group discussion yielded consensus on all items where there was not unanimous agreement among raters. Of the 40 items, 24 were scored as abstract, 11 were scored as concrete, three as formal, and two as primary. It is clear that there was not a great deal of difference in the orders of the items, which will be discussed further later in this paper.

A Rasch analysis was run on the items. This is a psychometric analysis of the responses to the items showing their relative difficulty (Rasch, 1980). The analysis yields two scales: the person's stage of performance and the Rasch-scaled item difficulty. A linear regression was then performed to compare the Rasch item difficulty with the prescored OHC of the items, with and without the word rarity variable. Then, Rasch person scores were regressed against participant Full Scale IQ (FSIQ) and the VCI.

The Rasch map (see Figure 2) indicated that the OHC scoring of items matched up fairly well with the item difficulty. Both of the primary items were at the very bottom of the map, and all of the concrete items were spread toward the very bottom. The most difficult item was one of the three formal items, and the other two were also among the most difficult. Meanwhile, the many abstract items were spread throughout the map, mostly occurring somewhere in the middle.

Regression Analysis

A simple linear regression was calculated to predict Rasch item scores based on the OHC of items. An $r = .666$ ($R^2 = .443$) was found, F(1,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Sample Scoring of Similarities Section

Points	Sample answers	In what way are a horse and a tiger alike?			Final verdict
		Rater 1	Rater 2	Rater 3	
2 points	Animals; mammals; members of the animal [kingdom, family]; quadrupeds; living things; alive	Abstract (shared variable classification)	Abstract (using abstract word to classify)	Abstract (recognizing the individual variables in the two animals, matching up the similar traits, and making a conclusion via classification)	Abstract
1 point	Both have [four legs, a tail]; have four legs and a tail [names shared physical feature(s)]; both are [powerful, strong, muscular, fast]; both can be tamed; can be pets; pets	Concrete (specific or concrete fact)	Concrete	Concrete (recognizing the individual variables in the two animals)	Concrete

38) = 30.275, $p < .001$. Rasch item difficulty is equal to −30.163 + 2.93 (OHC).

The rarity of an item word metric was included as a second independent variable in the next regression. Six vocabulary words (12 items) and one word from the similarities section (two items) were given a rarity score of 1, indicating rare words outside of the 10,000 most common words. The four abstract items with the highest Rasch difficulty were all rare words, indicating that rarity was an additional cause of difficulty. A multiple linear regression was calculated, where Rasch item difficulty = a_0 + a_1OHC + $a_2$rarity. An r = .778 was found, F(2, 37) = 28.831, $p < .001$; R^2 = .605. An item's Rasch difficulty score is equal to −25.199 + 2.323 (OHC) + 2.633 (rarity). The beta for OHC (.528) was higher than that of rarity (.425), as shown in Table 4. This is consistent with what was found in the factor analysis.

The next step of the analysis was to compare Rasch person scores to IQ scores. This would demonstrate the relationship between IQ and a loose estimate of stage. A regression of IQ (FSIQ) and Rasch person score (stage) had an r = .456, F(1, 99) = 27.435, $p < .001$; R^2 = .217. An additional linear regression had an r = .741 between Rasch person scores and the VCI, F(1, 99) = 120.456, $p < .001$; R^2 = .549.

Factor Analysis

An exploratory factor analysis was run on all of the items, including those that were not scored using OHC. Figure 3 is a scree plot of these data, demonstrating that the first factor had the largest eigenvalue and that there were a total of 49 factors found, most contributing almost nothing. The first factor had an eigenvalue of 7.834, explaining 15.67% of the variance. We posit that this factor is OHC. The second factor would be rarity, had an eigenvalue of 3.255, and explained an additional 6.51% of the variance. It is unknown what the third, fourth, and additional factors would be, and to attempt to understand this and support the argument that OHC and rarity were the first two items, the factor loadings of the individual items were looked at.

Table 5 lists all the items that loaded on each of the first four factors of more than .4, with the exception of the first factor, which only included items above .5. Many of the items with high factor

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

```
MEASURE   SUBJECT - MAP - RANK
    5    XXXXXXXXXX  +
                    T|
                     |
                     |T
                     |
                XXX  |
    4                +  fo1
                     |  ab22
                     |
          XXXXXXXXX  |
                    S|
                     |
    3       XXXXXXX  +
                     |  ab16
            XXXXXXX  |  ab14
                     |  ab20
                     |S
         XXXXXXXXXX  |
    2               M+
      XXXXXXXXXXXXX  |
          XXXXXXXXX  |  ab18   ab21   fo3
                     |
             XXXXXX  |  ab6
                     |  ab23   fo2
    1       XXXXXXX  +
                     |  ab8
                XXX  |
                    S|
            XXXXXXX  |
                     |  ab9
    0           XXX  +M
                     |  ab24
                     |  ab17
                     |  ab19
                XXX  |
                  X  |
   -1               T+
                  X  |
                     |  ab5
                  X  |  ab7
                     |  ab4
                  X  |
   -2                +
                     |  ab15   ab3    co10
                     |S
                     |  co4
                     |
                     |
   -3                +
                     |
                     |
                     |
                     |  ab11   ab13   co5
                     |
   -4                +  ab1    ab10   ab12   ab2    co1    co11   co2    co3
                        co6    co7    co8    co9    pr1    pr2
```

Figure 2. Rasch map of persons and items. The Xs to the left of the dashed line mark the person Rasch scores, and the items are on the right. The letter abbreviations mark order questions. The number marks the sequential number of that item among items of the same order. pr = primary; co = concrete; ab = abstract; fo = formal.

loadings were not used in scoring, and thus it is difficult to decipher too much from the table. However, it should be noted that the only similarities item that had a rarity score of 1 was the highest loading item on Component 2, supporting the argument that this is rarity. Additionally, the location of the information items helps support the idea that the first factor is OHC and the second factor is rarity. Because it was determined that the information section is mainly about knowing rare items and not stage, it makes sense that there are two items from the information section loading on the second factor and none on the first factor. Because all the items loading on the third and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Coefficients for Linear Regression Predicting Rasch Item Difficulty From OHC and Rarity

	Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
Model	*B*	*SE*	β	*t*	Sig.	Tolerance	VIF
(Constant)	−25.199	4.600		−5.478	.000		
OHC	2.323	.481	.528	4.834	.000	.895	1.118
Rarity	2.663	.683	.425	3.897	.000	.895	1.118

Note. Dependent variable is Rasch item difficulty. OHC = order of hierarchical complexity; Sig. = significance; VIF = Variance Inflation Factor.

fourth factors were unscored items, it is hard to say what these factors are.

A second exploratory factor analysis was run using only the items that were previously scored in order to help determine the third and fourth factors. This second-factor analysis again demonstrated that the first factor had a much larger eigenvalue than any other factor, as demonstrated in Figure 4, with this factor having an eigenvalue of 5.29, explaining 21.162% of the variance. There were a total of 25 factors, and we again extracted four and looked at the top item loadings. This second-factor analysis did not offer much support to our hypothesis that the first factor was OHC and the second was rarity (see Table 6). Most of the items loading on the first factor were scored as abstract, with a rarity of 1, indicating that rarity heavily loaded on this factor. The third component had only two factors with an eigenvalue above .5, and the fourth component had zero. From this information, it is difficult to determine what these factors may be. Additionally, it does not seem that these are highly relevant factors.

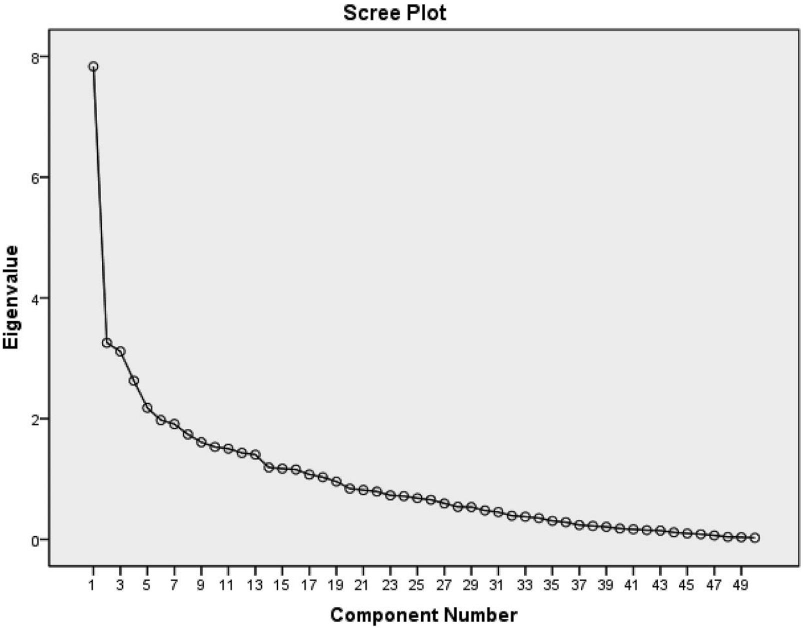

Figure 3. Scree plot of factors affecting all IQ items. The first component had by far the largest eigenvalue, and we propose that this is order of hierarchical complexity (OHC).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5
Eigenvalues of First Four Components of Exploratory Factor Analysis of All Items

Item	Eigenvalue	OHC	Rarity
Component 1			
Vocabulary 30	.657	—	—
Vocabulary 28	.626	Abstract	0
Similarities 18	.614		
Similarities 14	.607		
Vocabulary 27	.596	Abstract	0
Vocabulary 21	.560		
Vocabulary 19	.527	Abstract	1
Vocabulary 17	.527		
Vocabulary 23	.520		
Similarities 12	.512		
Vocabulary 16	.510		
Vocabulary 24	.500	Abstract	1
Component 2			
Similarities 7	.620	Abstract	
Information 16	.561		
Information 11	.441		
Similarities 9	.413	Abstract	
Component 3			
Information 18	.460		
Similarities 15	.404	Abstract	0
Component 4			
Vocabulary 8	.664		
Information 6	.501		
Vocabulary 14	.494		
Vocabulary 13	.413		

Note. OHC = order of hierarchical complexity.

Discussion

Just two variables—OHC and rarity—predicted Rasch IQ item difficulty with an r = .778. The variables were order of hierarchical complexity and rarity of the item. OHC was able to explain a great deal of the variance in the difficulty of the IQ items (r = .666). These results indicate that scoring items beforehand using the HCSS was successful. It demonstrates one of the key strengths of the MHC: that the difficulty of tasks can be assessed a priori. This is something that purely psychometric measures like IQ cannot claim to do because they do not know why certain questions are more or less difficult. This is a key finding that demonstrates the predictive ability of a psychophysical approach in attempting to understand "smarts."

Additionally, it became apparent in scoring the items that one of the factors that would affect the difficulty of the items would be how rare the words being tested were. The variable of rarity predicted the difficulty of items to a lesser degree than the OHC. This appears to be a flaw of the IQ test as a measure of ability because rarity was simply a measure of how likely the participants were to be exposed to those words. Undoubtedly, retention of knowledge plays a role in "smarts," but knowledge of rare words is extremely dependent on a participant's past history with a particular set of words. A person's knowledge of rare vocabulary words or facts is not something that will necessarily be a good predictor of success. Knowing rare things likely does not greatly affect a person's ability to succeed in jobs or other aspects of life and is often not what evaluators are trying to understand about a participant when assessing cognitive ability.

An additional finding of this study was that there was a correlation (r = .456) between Rasch person scores and IQ and an even greater correlation between Rasch person scores and verbal comprehension (r = .741). Rasch person scores are most likely a good representation of the stage scores of the participants. This indicates that there is a strong relation between stage and IQ.

It also must be noted that these stage scores are approximate. Due to the nature of the WAIS–IV, it is impossible to accurately determine the OHC of all the items. Therefore, what the true stages of the participants were was degraded. The fact that Rasch person scores were more closely correlated to the VCI (r = .741) is largely in part due to the fact that this was the section where the items tested were taken from. However, it also indicates that different sections measure slightly different constructs. This is something that the makers of the test acknowledge, but they believe that they measure verbal comprehension, working memory, perceptual reasoning, and processing speed, all of which factor into g. We argue that the VCI measures OHC, rarity, and some noise rather than g and verbal comprehension.

The factor analyses were run in an attempt to confirm that OHC and rarity were the two main factors; we also wanted to see if there were any other variables that we had missed. As discussed previously, using solely psychometric measures can lead to problems, but they are valuable tools. In this case, the factor

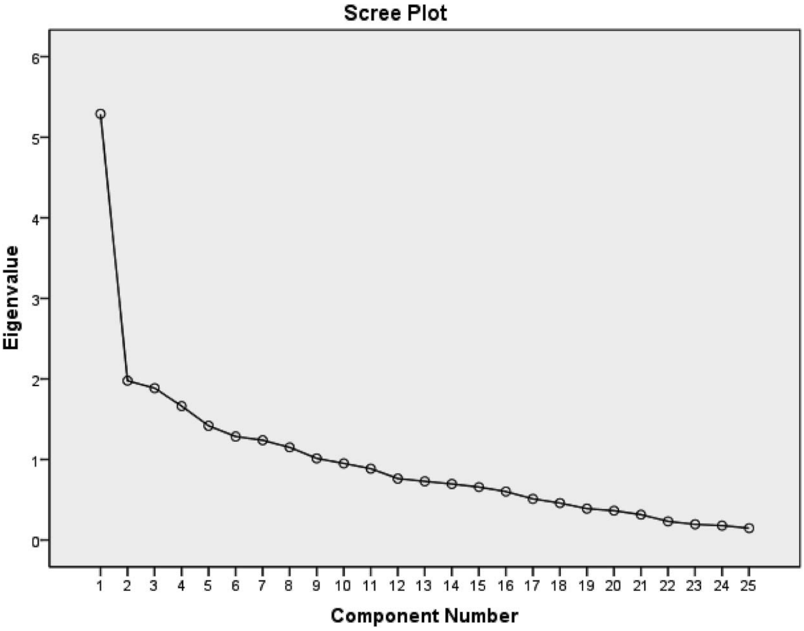

Figure 4. Scree plot of factors affecting only scored IQ items. The first component had by far the largest eigenvalue, and we propose that this is order of hierarchical complexity (OHC).

analyses demonstrated that there were two factors that contributed much of the variance but only provided limited support that those factors were OHC and rarity. The third and fourth factors had weak loadings and did not yield any obvious evidence of additional factors other than noise. Ultimately, because we could not score many of the items for OHC, the first-factor analysis yielded little information, and the second-factor analysis lost a lot of power by having to eliminate so many questions.

This difficulty in determining the stage of the participants largely comes from the lack of variance in the stage of the questions. Almost all of the items were scored as abstract or concrete. Meanwhile, the entire information section of the test asked the participants to recall facts, which was determined to be entirely concrete and was not used in the analysis. The lack of variability in stage is not a good way to design a test of cognitive ability, as ideally it would progress from lower to higher stage progressing all the way up to at least the metasystematic stage. If several questions are incorrectly answered in a row, the WAIS–IV stops the participant from answering further questions. However, because it is not clear why the more difficult items are more difficult for the participant, stopping at a certain point does not give any information as to what the participant has failed at. If the test were designed with sequentially more complex items, it would be possible to see what order of complexity a participant was struggling with. Instead, the IQ test yields an arbitrary number based on norms as an IQ score. This does not give any insight to academic instructors, clinicians, employers, or whoever is administering the test about what specific tasks a participant may struggle with or succeed at. From a complexity perspective, all that can really be gathered is whether an individual is able to successfully operate at the abstract order.

In addition to the simple lack of variance and correct ordering of questions, there were only three formal order questions and nothing higher. This means that it was not possible to give an accurate prediction of stage beyond the formal stage and not even truly accurate assessments of whether a participant reached the formal stage. This is a major flaw of the IQ test because there are several stages be-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 6
Eigenvalues of First Four Components of Exploratory Factor Analysis of Scored Items

Item	Eigenvalue	OHC	Rarity
Component 1			
Vocabulary 29b	.697	Abstract	1
Vocabulary 28b	.661	Abstract	1
Vocabulary 27b	.635	Abstract	1
Vocabulary 27a	.605	Abstract	1
Vocabulary 29a	.602	Formal	1
Vocabulary 26a	.546	Abstract	1
Similarities 9a	.543	Abstract	0
Vocabulary 28a	.527	Abstract	1
Vocabulary 24b	.520	Abstract	1
Component 2			
Similarities 6b	.710	Abstract	0
Similarities 6a	.638	Abstract	0
Similarities 7a	.610	Abstract	1
Component 3			
Vocabulary 15b	.658	Abstract	0
Similarities 5a	.546	Abstract	0
Component 4			
Similarities 17b	.471	Abstract	0
Vocabulary 7a	.432	Concrete	0
Vocabulary 15b	.432	Abstract	0
Vocabulary 19b	.416	Abstract	1

Note. OHC = order of hierarchical complexity.

yond the formal stage that adult humans are very capable of reaching, and yet this test operates as if formal stage is the highest stage that exists. Commons, Miller, and Giri (2014) report that 20% of educated adults reach Systematic Stage 12 (one stage beyond Formal Stage 11), with an additional 1.5% reaching Metasystematic Stage 13.

Limitations and Difficulties

One difficulty that emerged during this study was that the IQ items were not designed with OHC in mind, and there were thus many difficulties scoring the items. The 2-point sample answers for a single item might vary widely in stage, making it impossible to determine what stage a participant actually answered at. This is a problem with the way the IQ test is designed because a more complex answer is not necessarily rewarded properly. However, it also meant that the answers could not be scored entirely accurately. An ideal study would record the individual's answers as they responded and then score their stage; however, this would be a more arduous task. Perhaps this could be addressed in future studies.

An additional limitation of this study was that the participants were all "high IQ," further decreasing the variance. This was a convenience sample, as the participants were being given the IQ test as part of their normal assessment; the goal of the authors was not to solely evaluate high-IQ participants. The mean IQ was 121.34, well above the population mean of 100. Four participants got the full 2 points on every question we analyzed, and several more got close-to-perfect scores. There were also a few questions every participant received 2 points on, essentially rendering the question useless. This problem is partially due to the fact that the IQ test is based on norms and does not have a very good measure of people at the very high and low ends of the spectrum. The MHC does not rely on norms and so would not have these problems.

Furthermore, in this study, only those questions that were identical to the English version of the WAIS–IV were used in the analysis. This was in order to decrease errors in the translation of the scoring manual but decreased the number of items that could be used. This similarly decreased the variance, and with more items the *r* could have been higher. The fact that the French and English versions had many differences is also an interesting characteristic of the WAIS–IV. If the translation is not the same, the test may be testing different constructs in different cultures because, as discussed previously, the creators do not have a good idea of what exactly they are testing. This study also only used questions from the VCI and not the entire test. This means that any generalizations to the entire WAIS–IV and the measurement of IQ in general cannot strictly be made from this study alone. However, the other sections of the IQ test were judged to not be easily scored in terms of stage, which indicates that they do not assess the solving of complex tasks.

Conclusion and Future Directions

As discussed, this study had a sample of French-speaking, above-average-IQ participants. Future studies would benefit from a more diverse sample that would increase the variance.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

While it is unlikely that culture impacts stage (Giri, in press), and the WAIS–IV translations were psychometrically designed to have the same distribution across different languages, studies from different cultures and languages could prove useful. An ideal study for the future would consist of recording participants' responses to the items directly and then scoring them according to the HCSS. These scores could then be used to get a more accurate measure of stage, which could then be compared to the participant's IQ score.

There are hundreds of studies with good instruments on development. However, predictive studies like those that are common with IQ are rare, and often the methodology is weak. Therefore, what the future portends is applying both stage and IQ at the same time and comparing how well each factor predicts success—for example, predicting job performance using both IQ and stage.

This study demonstrates the ceiling effect of the VCI of the WAIS–IV as well as its shortcomings in assessing solving complex tasks. The order of hierarchical complexity predicted which items would be most difficult in conjunction with an irrelevant variable of word rarity. This demonstrates how the OHC is a successful indicator of task difficulty and supports the argument that a person's stage, as measured by their ability to successfully complete tasks of a certain order, is a strong indicator of intelligence.

However, even though IQ tests are only relative measuring instruments and have their limitations, they are currently the most common and easiest clinical tools to give a psychometric value to the intellectual abilities of a person. The combination of IQ with clinical history and other elements enables clinicians to obtain a psychological and cognitive profile of the person, which can be important for clinical follow-ups. While assessments of stage based on the MHC exist, these tools need to be further refined and assessed in order to develop an assessment that can be easily used and understood in the daily practice of psychologists.

References

Binet, A., & Simon, T. (1914). *Mentally defective children*. London, United Kingdom: Longmans, Green.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. http://dx.doi.org/10.1007/s11336-006-1447-6

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The model of hierarchical complexity as a measurement system. *Behavioral Development Bulletin, 19,* 9–14.

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K., & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement, 9,* 182–199.

Commons, M. L., Miller, L. S., & Giri, S. (2014). A model of stage change explains the average rate of stage of development and its relationship to the predicted average stage ("smarts"). *Behavioral Development Bulletin, 19,* 1–11. http://dx.doi.org/10.1037/h0101076

Commons, M. L., & Miller, P. M. (1998). A quantitative behavior-analytic theory of development. *Revista Mexicana de Análisis de la Conducta, 24,* 153–180.

Commons, M. L., Miller, P. M., Goodheart, E. A., & Danaher-Gilpin, D. (2005). *Hierarchical Complexity Scoring System: How to score anything*. Retrieved from http://www.dareassociation.org/Papers/Scoring%20Manual.htm

Commons, M. L., & Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures: The Journal of New Paradigm Research, 64*(5–7), 375–382. http://dx.doi.org/10.1080/02604020802301204

Commons, M. L., & Richards, F. A. (1984). Applying the general stage model. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development* (pp. 141–157). New York, NY: Praeger.

Commons, M. L., Richards, F. A., & Kuhn, D. (1982). Systematic and metasystematic reasoning: A case for a level of reasoning beyond Piaget's formal operations. *Child Development, 53,* 1058–1069. http://dx.doi.org/10.2307/1129147

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review, 8,* 237–278. http://dx.doi.org/10.1006/drev.1998.0467

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience, 11,* 201–211.

DeVries, R. (1974). Relationships among Piagetian, IQ, and achievement assessments. *Child Development, 45,* 746–756. http://dx.doi.org/10.2307/1127841

Dudek, S. Z., Lester, E. P., Goldberg, J. S., & Dyer, G. B. (1969). Relationship of Piaget measures to standard intelligence and motor scales. *Perceptual and Motor Skills, 28,* 351–362. http://dx.doi.org/10.2466/pms.1969.28.2.351

Giri, S. (in press). Crosscultural homogeneity in social perspective taking. *Behavioral Development Bulletin*.

Humphreys, L. G., & Parsons, C. K. (1979). Piagetian tasks measure intelligence and intelligence tests assess cognitive development: A reanalysis. *Intelligence, 3,* 369–381. http://dx.doi.org/10.1016/0160-2896(79)90005-9

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the development of formal operational structures* (A. Parsons & S. Seagrim, Trans.). New York, NY: Basic Books. http://dx.doi.org/10.1037/10034-000 (Original work published 1955)

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2005). A history of intelligence test interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 23–38). New York, NY: Guilford Press.

Kohlberg, L., & Mayer, R. (1972). Development as the aim of education. *Harvard Educational Review, 42,* 449–496. http://dx.doi.org/10.17763/haer.42.4.kj6q8743r3j00j60

Mackintosh, N. J. (2011). *IQ and human intelligence* (2nd ed.). Oxford, England: Oxford University.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28,* 1–14. http://dx.doi.org/10.1037/h0034092

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

Pelli, D. G., & Farell, B. (1995). Psychophysical methods. In M. Bass, E. W. Van Stryland, D. R. Williams, & W. L. Wolfe (Eds.), *Handbook of optics* (2nd ed., Vol. I, pp. 29.21–29.13). New York, NY: McGraw-Hill.

Rasch, G. (1980). *Probabilistic model for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Schönemann, P. H. (1997). On models and muddles of heritability. *Genetica, 99*(2–3), 97–108. http://dx.doi.org/10.1007/BF02259513

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV)*. San Antonio, TX: Psychological Corporation.

White, S. (2000). Conceptual foundations of IQ testing. *Psychology, Public Policy, and Law, 6,* 33–43. http://dx.doi.org/10.1037/1076-8971.6.1.33

Received September 21, 2015
Revision received March 9, 2016
Accepted March 14, 2016 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.