Evaluating within CHC domain test score differences

Deciding when the scores from two tests, which are from the same CHC domain (e.g., *Gwm*), and which may have the same narrow CHC classifications, are different enough to warrant clinical interpretation.

Understanding score exchangeability

It is important to understand the correlations between similarly classified CHC tests and not assume that because they are significantly correlated (or load on the same CHC factor in factor analysis) their scores are interchangeable. Divergent scores within many CHC domains will occur with regularity.

The coefficient of determination (correlation squared X 100) provides the most important information regarding the shared variance between tests.

For example. Two *Gwm* tests that correlate .60 share approximately 36% common variance (.60 x .60 X 100). Although a moderate and significant correlation, the scores from these two *Gwm* tests actually have more that they don't share in common (64% variance divided up into error variance and unique unshared variance), than they have in common (36% shared score variance)

Select WJ IV COG and WISC-IV similarly CHC-classified tests correlations (n=173)—Study in WJ IV technical manual (McGrew, et al., 2014)

	Vocab	Info	Digit Span	Let-Nm Seq	Coding	Sym Search	Canc.	Block Design
Oral Vocabulary	0.78							
General Information		0.68						
Verbal Attention			0.58	0.53				
Numbers Reversed			0.65	0.45				
Object-Number Sequencing			0.45	0.42				
Letter-Pattern Matching					0.42	0.37	0.36	
Number Pattern-Matching					0.50	0.48	0.44	
Visualization								0.53

46% to 61% shared variance

18% to 25% shared variance

28% shared variance

18% to 42% shared variance

Similarly CHC narrow classified tests (within and across batteries) should not be assumed to be 1-1 exchangeable

School Psychology Review,
2005, Volume 34, No. 3, pp. 329-357

Are Cattell–Horn–Carroll Broad Ability Composite Scores Exchangeable Across Batteries?

Randy G. Floyd
Renee Bergeron
Allison C. McCormack
Janice L. Anderson
Gabrielle L. Hargrove-Owens
The University of Memphis

Abstract. Many school psychologists use the Cattell–Horn–Carroll (CHC) theory of cognitive abilities to guide their interpretation of scores from intelligence test batteries. Some may frequently assume that composite scores purported to measure the same CHC broad abilities should be relatively similar for individuals no matter what subtests or batteries were administered to obtain these scores. This study examined this assumption using six samples of preschool children, school-age children, or adults who completed two or more intelligence test batteries. From these samples, composites measuring the broad abilities Crystallized Intelligence, Visual Processing, Fluid Reasoning, and Processing Speed were compared to examine their exchangeability. Results indicate that most CHC broad ability composites produced scores that were not as exchangeable for individuals as may have been assumed by some. Discussion focuses on the influence of score reliability and on the interaction between examinee characteristics and the tasks used to measure the broad abilities.

Important article to read regarding score exchangeability

Correlations between WJ IV *Gwm* tests and tests with possible *Gwm* variance (based on WJ IV norm subjects from ages 6-19)

Pearson Correlation Matrix

	VRBATN	NUMREV	OBJNUM	UNDDIR	MEMWRD	SENREP	NWDREP	STYREC	RDGREC
VRBATN	1.00								
NUMREV	0.47	1.00							
OBJNUM	0.56	0.48	1.00						
UNDDIR	0.46	0.33	0.43	1.00					
MEMWRD	0.51	0.43	0.50	0.37	1.00				
SENREP	0.48	0.29	0.39	0.49	0.49	1.00			
NWDREP	0.45	0.28	0.45	0.45	0.43	0.50	1.00		
STYREC	0.34	0.29	0.42	0.42	0.37	0.28	0.33	1.00	
RDGREC	0.34	0.28	0.27	0.33	0.18	0.28	0.22	0.32	1.00

Shared variance among primary WJ IV *Gwm* tests ranges from 08% to 31%.

Scores from these WJ IV *Gwm* tests are not interchangeable and divergent scores are going to occur with regularity .

Select WAIS-IV *Gwm* (working memory) correlations—all ages (Table 5.1; p 62 WAIS-IV technical manual)

	DS	AR	LNS
Digit Span (DS)	--		
Arithmetic (AR)	**.60**	**--**	
Letter-Num. Sequencing (LNS)	**.69**	**.56**	--

Shared variance among WAIS-IV *Gwm* tests ranges from 31% to 48%.

	DSF	DSB	DSS
Digit Span Forward (DSF)	--		
Digit Span Backward (DSB)	**.58**	**--**	
Digit Span Sequencing (DSS)	**.42**	**.51**	--

Shared variance among WAIS-IV Digit span tests ranges from 18% to 34%.

Scores from these WAIS-IV *Gwm* tests are not interchangeable and divergent scores are likely to occur with regularity.

Understanding score difference base rates

Before interpreting differences between two similarly classified CHC tests (e.g., two *Gwm* tests), it is important to first determine if the difference is significant and unusual.

STATISTICS FOR THE INVESTIGATION OF INDIVIDUAL CASES*

R. W. PAYNE AND H. GWYNNE JONES

Institute of Psychiatry, University of London, Maudsley Hospital

PROBLEM

Much of the work of a clinical psychologist consists of making relatively routine psychological measurements of fairly well established traits, either cognitive or orectic. It is well known, however, that there can be no measurement without error. The psychologist must have the means of taking error into account if he is to assess his test scores intelligently. There appear to be three main types of question which face clinical psychologists:

1. *The Abnormality of a Discrepancy between Two Scores*

This problem arises every time a psychologist gives more than one measure. Perhaps the commonest example is the Wechsler-Bellevue Intelligence Scale. This test provides two rather different measures of intelligence, the "Verbal Scale IQ" and the "Performance Scale IQ". It is a common experience that these two scores are divergent. In fact the discrepancy may suggest interesting hypotheses in line with other abnormalities the patient shows. However, before we can assess such a discrepancy, we must take into account two factors. We know that neither scale is perfectly reliable and we also know that the scales are not perfectly correlated. Therefore, many normal people would show discrepancies between the two scales which one need not take seriously. The first question we can ask ourselves then, is how frequently would a discrepancy as large as the one we observe occur in the normal population? That is, how "abnormal" is the difference we observe between our test scores?

2. *The reliability of a discrepancy between two scores*

In certain cases, we may have occasion to give two tests which measure rather different traits. For example, we may give a test of long term retention, and a test of general intelligence. It may be the case that these tests have a very low intercorrelation in the general population, so that quite large discrepancies between these scores could be quite "normal" or usual in the general population. Nevertheless on clinical grounds, we might expect our patient to have a lower memory test score than a general intelligence test score. We are not implying that this would be an abnormally *large* discrepancy. Many people may have as large differences. We *are* implying, however, that it is a *measurable* difference. We know that neither test is perfectly reliable, so that small differences will occur by "chance". What we wish to know is how large a difference between any two scores must be before we can be sure the difference could not be due merely to error of measurement of the tests.

3. *Testing a Clinical Prediction*

A third type of problem is slightly different. Very often the clinical psychologist finds himself repeating a measurement with a certain expectation or "prediction". For example, a patient may obtain an "average" IQ when first seen. Two years later, there may be strong clinical grounds for believing that deterioration has taken place. We, therefore, wish to retest him on the same (or a similar) test of intelligence to confirm the hypothesis that he has deteriorated. We may, indeed, find that his score is now below average. Have we in fact confirmed our hypothesis?

Again we know that tests are not perfectly reliable and that such changes in score occur in perfectly normal people. Essentially we need a control group. We need to know what proportion of individuals like our patient, of the same IQ on the first

It's a pleasure when you use the correct measure

Three primary models for evaluating score differences

(Payne & Jones, 1957)

www.iapsych.com/articles/payne1957.pdf

It's a pleasure when you use the correct measure

Three primary models for evaluating score differences

A. *Evaluating "abnormality" (base rate) of a difference score* (Payne & Jones, 1957). If difference is a simple difference score, and the explicit emphasis is on the cohesiveness (correlation) of tests within a composite/CHC domain, then the *SD(diff)* is a better statistic.

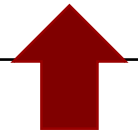
B. *Evaluating the reliability of a difference score* (Payne & Jones, 1957). If the difference is a simple difference score, and the tests measure rather different traits (e.g., not within same broad CHC domain; low correlation/cohesion), then one can use the reliability of difference scores—*SE(diff).*

C. *Evaluating a prediction* (Payne & Jones, 1957). If the difference implies a predictive relationship, then regression to the mean needs to be accounted for and the proper statistic is the *SE(est).*

Reliability (Is there a difference?) vs. Abnormality (How unusual is the difference?)

(Distinction and table courtesy of Dr. Joel Schneider)

	Simple Difference (X – Y)	**Prediction Error (Y – Ŷ)**
Reliability	Are these 2 scores different?	Is this outcome different from expectations?
Abnormality (base rate)	How unusual is it for these 2 scores to differ by this much?	How unusual is it for this outcome to differ from expectations by this much ?

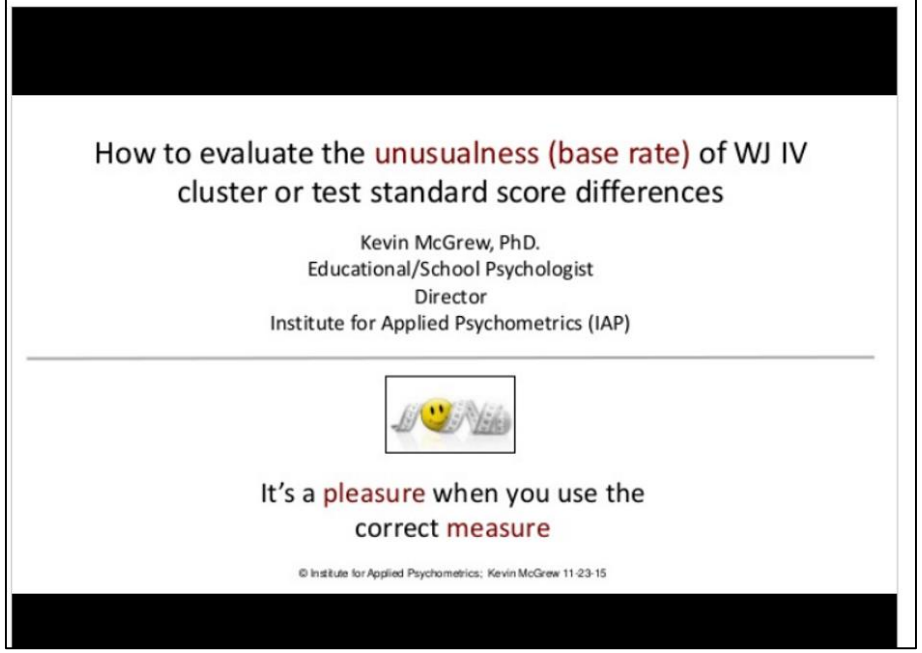


This is the most important issue when determining if scores from two tests within the same CHC domain (e.g., *Gwm*) are discrepant enough to warrant interpretation of the difference. Often called evaluating the "cohesion" of scores within a CHC domain

How to evaluate the unusualness (base rate) of WJ IV cluster or test score differences: It is a pleasure to use the correct measure - A SlideShare presentation

The WJ IV provides two primary methods for comparing tests or cluster scores. One is based on a predictive model (the variation and comparison procedures) and the other allows comparisons of SEM confidence bands, which takes into account each measures reliability. A third method for comparing scores, one that takes into account the correlation between compared measures (ability cohesion model) is not provided, but is frequently used by assessment professionals. The three types of score comparison methods are described and new information, via a "rule of thumb" summary slide and nomograph, are provided to allow WJ IV users to evaluate scores via all three methods.

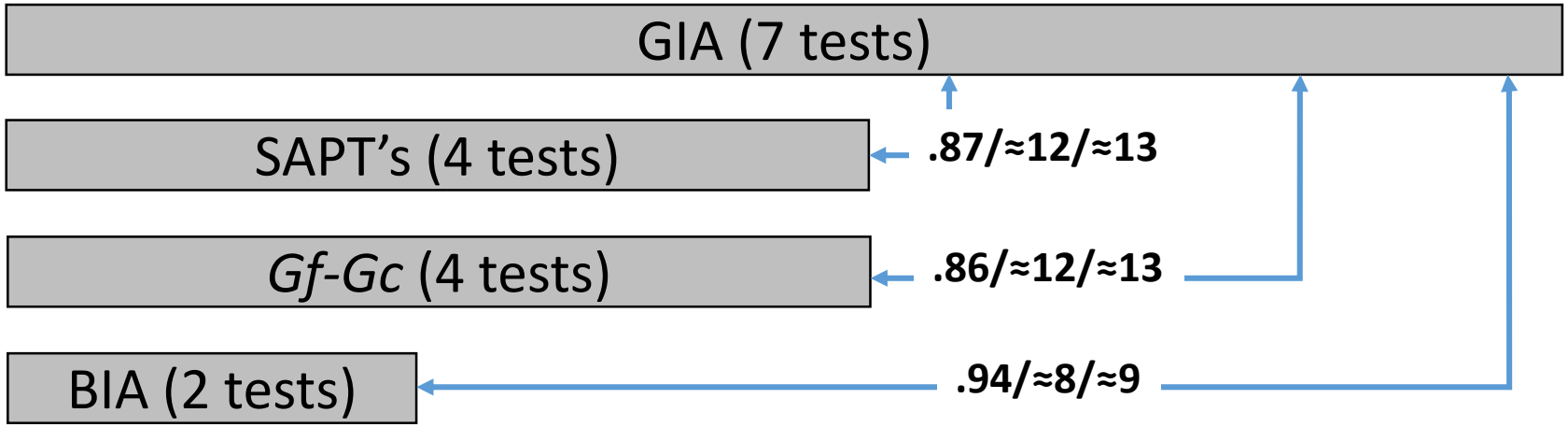
A PDF copy of the key WJ IV base rate rule-of-thumb slide can be found here.



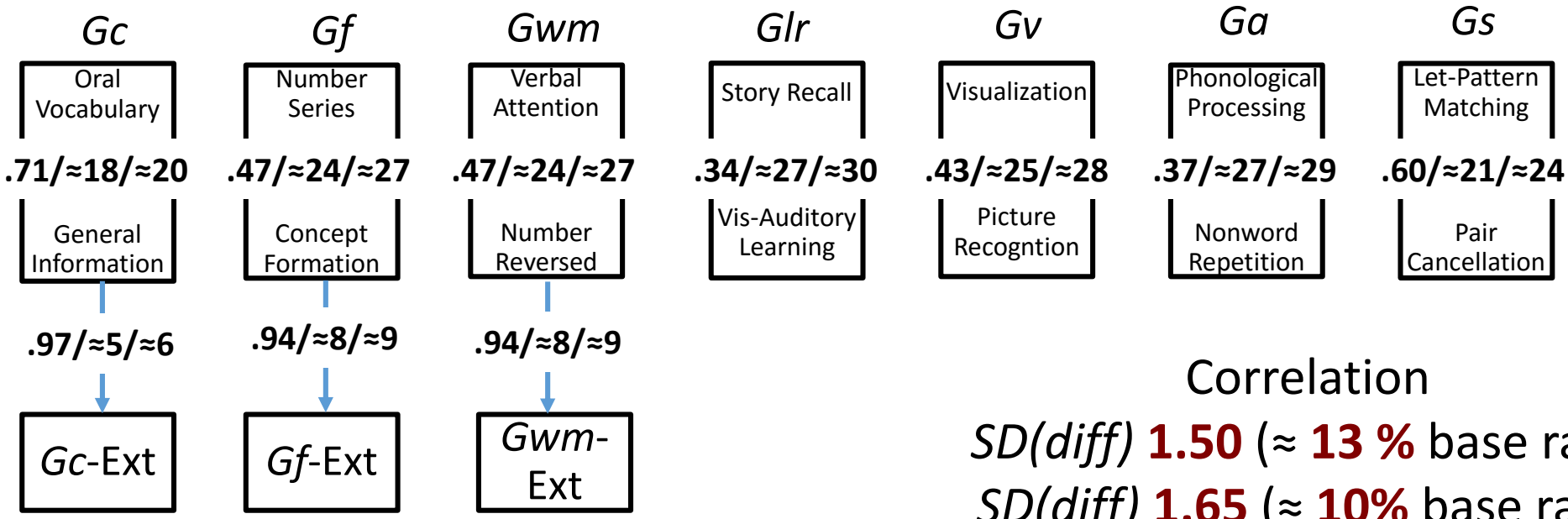
Visit IQ's Corner for a more detailed slide show explanation

http://www.iqscorner.com/2016/02/how-to-evaluate-unusualness-base-rate.html

Select WJ IV COG cluster/test score significance values (ages 6-19) *



* Rounded values calculated in WJ IV norm data (ages 6 to 19)



Correlation

SD(diff) **1.50** (≈ **13 %** base rate)

SD(diff) **1.65** (≈ **10%** base rate)

How to use the information on prior slide

Gwm

Verbal Attention

.47/≈24/≈27

Number Reversed

WJ IV Verbal Attention and Numbers Reversed tests correlate, on average, at .47. This indicates approximately 27% shared score variance—they are not interchangeable.

A SS difference of 24 points or more is needed to be unusual at 1.50 SD(diff) – 13% base rate

A SS difference of 27 points or more is needed to be unusual at 1.65 SD(diff) – 10% base rate

It is highly recommended, when using the WJ IV battery, to pay even more attention to the relative performance index (RPI) scores for two tests being compared, and less attention the SS differences.

Tests with similar SS's can have markedly different RPI's

The RPI provides a "real world" functional metric that better describes how the person tested is likely to perform on similar tasks—"where the rubber meets the road."

Example from WJ III case

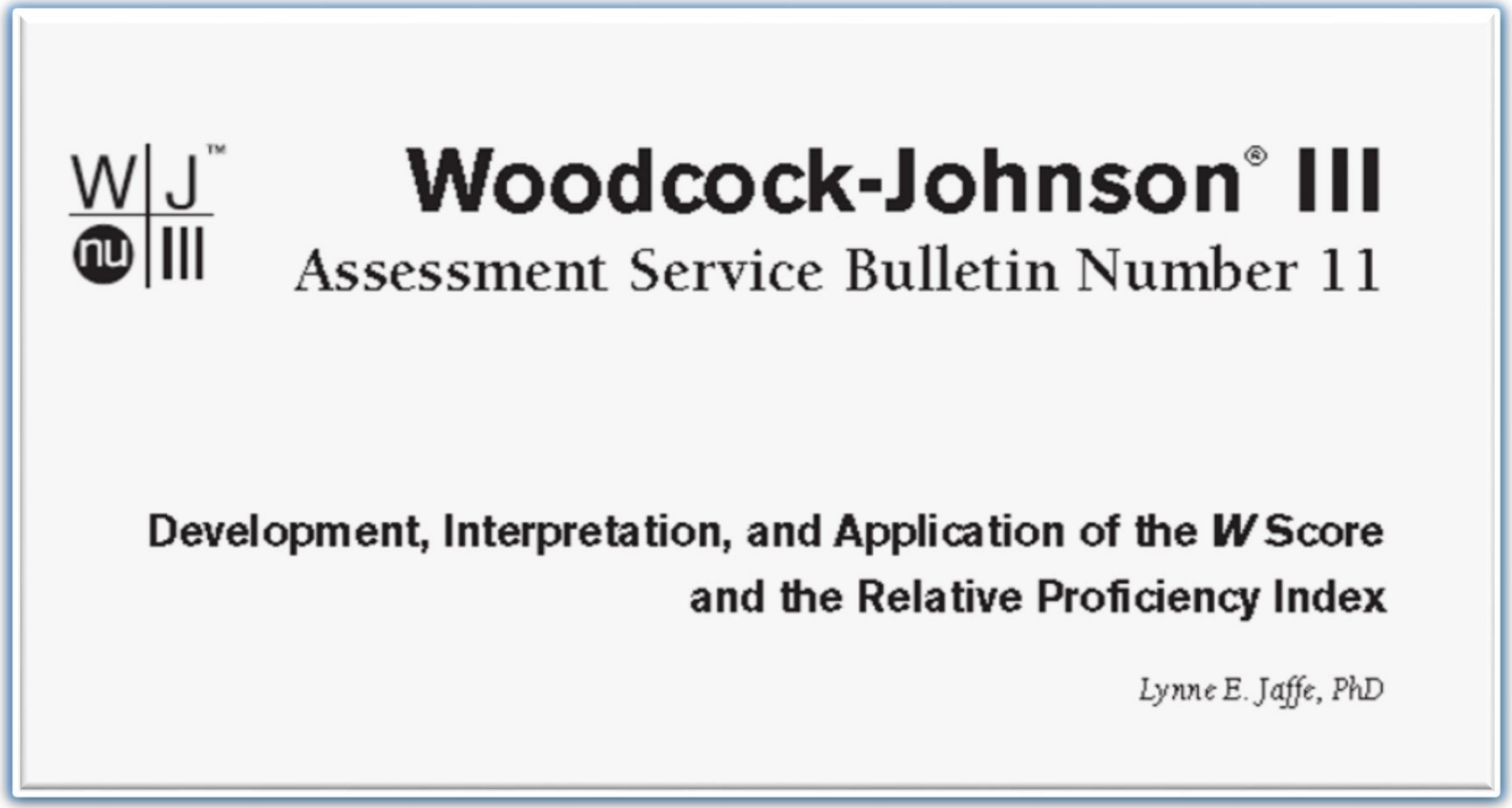
- **12 year old**
- **Numbers Reversed SS = 86 RPI = 48/90**
- **Aud. Work. Memory SS = 88 RPI = 69/90**

This individual is expected to perform with 48% mastery or proficiency on these type of cognitive when others of the same age/grade perform with 90% mastery or proficiency

VS

This individual is expected to perform with 69% mastery or proficiency on these type of cognitive when others of the same age/grade perform with 90% mastery or proficiency

The reality of expected level of mastery or proficiency is not reflected in the 2 SS point difference but is clear when comparing the RPI's



A must read

http://www.hmhco.com/hmh-assessments/cognitive-intelligence/wj-iii-nu#assessment-service-bulletins

More to come on this topic. Stay tuned to IQ's Corner blog and the IAP CHC listserv