Aggregation of AC Scores
A Commentary on Recent Perspectives

Presented at: 38th International Congress on Assessment Center Methods
Alexandria, VA

Presenter: Dan J. Putka, HumRRO

October 23, 2014
Introduction

- Much research has examined the latent structure of assessment center (AC) ratings

- Research has largely focused on within-exercise dimension ratings (WEDRs)

- Until recently, few have considered how aggregation of WEDRs can fundamentally alter conclusions regarding the latent structure, and in turn, validity and reliability of AC scores

- But wait…in practice we tend to use WEDRs that have been aggregated to form dimension-level or overall AC scores?
  - Disconnect between research focus of past 30 years and practice!
Two recent JAP articles

- **Putka & Hoffman (2013)**
  - Aggregation, *and* the generalizations one wishes to make regarding individuals’ scores fundamentally affect the *proportion and composition* of “reliable” and “unreliable” variance in AC scores
    - Will they generalize to new assessors, new exercises, etc.?
    - It makes no sense to talk about “the” reliability of AC scores
  - Application of random effects models to WEDR data

- **Kuncel & Sackett (2014)**
  - Aggregation fundamentally affects the *proportion* of “construct relevant” and “construct irrelevant” variance in AC scores
  - Application of algebra of covariances to CFA models of WEDR data
Bottom line

- Using very different methods (random effects models vs. CFA) and frames of reference (reliability vs. validity) these studies show that as WEDRs are aggregated, the relative contribution of assesssee, dimension, exercise-related effects to observed variance in AC scores can change substantially.

- This has implications for what one concludes regarding the reliability and validity of AC scores.

- This also suggests that AC scores that may appear unreliable or invalid at one level of aggregation may be perfectly reasonable at another level.
So what?
Why we should care

- Collectively, these studies have real implications for how we judge the psychometric quality of our AC scores and interpret past research – an issue for AC research AND practice

- Aggregation definitely matters numerically, but there are additional *fundamental issues* that undergird the *interpretation* of these studies’ results that may get lost in the shuffle

- As a set, these two studies provide an excellent context for illustrating these fundamental issues and improving the way our field approaches discussing and evaluating the reliability and validity of AC scores
We often draw firewalls between validity and reliability, but these concepts are *layers of meaning* we ascribe to variance decomposition results and analytic techniques which are “blind” to that layer of meaning.

The two articles find similar things with regard to distribution of variance in AC scores and role of aggregation, but the *layer of meaning* attached to them differs.
Fundamental issue #2

- The *same* source of variance can legitimately be viewed as a source of reliable or unreliable variance depending on...
  - One’s AC *measurement design*
  - One’s *desired generalizations*
  - One’s *use of AC scores*
  - The underlying distribution of variance and effect of aggregation doesn’t change, *but the layer of meaning* applied to it does

- This has long been appreciated in the educational measurement arena, but slower to catch on in I-O
  - Putka & Hoffman, 2013 – ACs
  - Putka & Hoffman, 2014 – performance ratings
  - Putka & Sackett, 2010 – selection measures in general
Fundamental issue #3

- The *same* source of variance can legitimately be viewed as a source of construct relevant or construct irrelevant variance depending on...
  - How one *designed his/her AC exercises*
    - Did you intend your exercises to provide *parallel measures* of a given dimension?
  - One’s *measurement and evaluation paradigm*
    - Traditional individual differences paradigm
    - Interactionist paradigm
  - The underlying distribution of variance and effect of aggregation doesn’t change, *but the layer of meaning* applied to it does

- Arguably, this has yet to be widely appreciated in ed. measurement or I-O
  - K&S (2014) frame their article, results, and conclusions from the perspective of only ONE measurement and evaluation paradigm
Today’s talk

- **Objective 1**: Visualize variance in AC scores
- **Objective 2**: Visualize the effects of aggregation
- **Objective 3**: Discuss the layer of meaning we ascribe when tying variance partitioning results to issues of reliability and validity
- **Objective 4**: Discuss the implications of these observations for evaluating the quality of AC scores in research and practice
Visualizing Variance in AC Scores

Pie Chart Paradise
Partitioning variance

- Regardless of whether we set out to investigate construct validity or reliability issues, AC researchers are concerned with the *latent* composition of observed variance in AC scores.

- What do we mean by latent?
  - Unseen
  - In the context of validity investigations, we typically associate “construct relevant variance” and “construct irrelevant variance” with *latent factors*.
  - In the context of reliability investigations, we typically associate “true score variance” and “error variance” with one or more latent, orthogonal *variance components*.
**What we see**

**Observed variance** between candidates on a given dimension as assessed within a single exercise (WEDR)

<table>
<thead>
<tr>
<th>Person</th>
<th>Score on Dimension1 - Exercise 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
What we estimate

**Unobserved components of variance** between candidates on a given dimension as assessed within a single exercise (WEDR)
What we estimate

Random Effects Model: person $\times$ dimension $\times$ exercise $(p \times d \times e)$

Random effects models underlie Generalizability theory
What we estimate

1 General Factor - Uncorrelated Dimensions – Uncorrelated Exercises
CFA Model (1G-UD-UE)
What we estimate

Correlated Dimensions – Correlated Exercises CFA Model (CD-CE)
What we can estimate...with multiple assessors

Random Effects Model: person × dimension × exercise × assessor (p × d × e × a)

Sources of Consistency Across Assessors
What we can estimate... with multiple assessors

Random Effects Model: person × dimension × exercise × assessor
\((p \times d \times e \times a)\)

Sources of **Inconsistency** Across Assessors
Visualizing the Effects of Aggregation

More Pie Chart Fun
Effects of aggregation

Random Effects Model: person × dimension × exercise (p × d × e)

Averaging WEDRs to form Dimension-Level AC Scores

Example of averaging across two exercises.
Effects of aggregation

Random Effects Model: person × dimension × exercise (p × d × e)

Averaging WEDRs to form Overall AC Scores

Example of aggregating across two exercises, each measuring the same two dimensions.
Random Effects Model: person $\times$ dimension $\times$ exercise $\times$ assessor ($p \times d \times e \times a$)

WEDRs Before Aggregation

Dimension-Level Scores: Aggregating WEDRs Across 3 Exercises

$\Delta = \text{Aggregated component } \% / \text{Unaggregated component } \%$
Correlated Dimensions – Correlated Exercises CFA Model (CD-CM)

Let each component (d.g, d.s, etc.) equal $Y$.

Solve the following equation for each component $Y$, where $X_i$ equal scores on a dimension as measured within exercise $(i)$:

$$rY\sum_{1}^{k} X_i = \frac{\sum CovYX_i}{(\sum Var(X_i) + 2\sum \sum CovX_iX_j)^{1/2}}$$

Square the resulting correlation and we have the proportion of variance in *dimension-level* scores attributable to the component.
Kuncel & Sackett (2014)

Correlated Dimensions – Correlated Exercises CFA Model (CD-CM)

WEDRs Before Aggregation

Dimension-Level Scores: Aggregating WEDRs Across 3 Exercises

Based on analysis of Bowler & Woehr (2006) meta-analytic data

\[ \Delta = \frac{\text{Aggregated component \%}}{\text{Unaggregated component \%}} \]

Dimension-Level Scores: Aggregating WEDRs Across 3 Exercises

P&H (2013)

K&S (2014)

\[ \Delta = \text{Aggregated component %} / \text{Unaggregated component %} \]
#1: Overall, aggregation impacts components similarly across studies, and the relative distribution of variance is comparable, but there are some key differences…

Dimension-Level Scores: Aggregating WEDRs Across 3 Exercises


#2: K&S report a bump in d.s variance, but that effect confounds dimension-specific variance that is consistent across assessors (pd) and dimension-specific variance that is inconsistent across assessors (pda).
#3: K&S report a large “residual error” variance component, but a good portion of that appears to be systematic dimension-exercise interactions (meaningful from a interactionist – trait activation theory perspective).

Dimension-Level Scores: Aggregating WEDRs Across 3 Exercises

**P&H (2013)**

- pde
- pe
- pda

**Source**

<table>
<thead>
<tr>
<th>Source</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.35</td>
</tr>
<tr>
<td>pa</td>
<td>0.02</td>
</tr>
<tr>
<td>pd</td>
<td>0.02</td>
</tr>
<tr>
<td>pda</td>
<td>0.12</td>
</tr>
<tr>
<td>pe</td>
<td>0.24</td>
</tr>
<tr>
<td>pea</td>
<td>0.01</td>
</tr>
<tr>
<td>pde</td>
<td>0.16</td>
</tr>
<tr>
<td>r</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**K&S (2014)**

- d.g
- d.s
- e.s
- e.g

**Source**

<table>
<thead>
<tr>
<th>Source</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.g</td>
<td>0.35</td>
</tr>
<tr>
<td>d.s</td>
<td>0.09</td>
</tr>
<tr>
<td>e.g</td>
<td>0.07</td>
</tr>
<tr>
<td>e.s</td>
<td>0.20</td>
</tr>
<tr>
<td>r</td>
<td>0.29</td>
</tr>
</tbody>
</table>

**#4**: P&H “collapse” d.g. and e.g into a general person main effect, but its magnitude is fairly similar to K&S’s d.g and e.g effects combined. Is it meaningful to have dim. factors uncorrelated with ex. factors (the case in CD-CE models)?
Layers of Meaning: Reliability & Validity

Sorry, No More Pie Charts
Putka & Hoffman (2013) - Reliability

Reliability Depends on Aggregation…

<table>
<thead>
<tr>
<th>Level of Score</th>
<th>WEDR</th>
<th>Dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.18</td>
<td>0.35</td>
</tr>
<tr>
<td>pa</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>pd</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>pda</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>pe</td>
<td>0.36</td>
<td>0.24</td>
</tr>
<tr>
<td>pea</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>pde</td>
<td>0.24</td>
<td>0.16</td>
</tr>
<tr>
<td>r</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>rxx</td>
<td>0.19</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Dimension level scores reflect averaging 3 WEDRs across exercises

Reliability Depends on Desired Generalizations…

<table>
<thead>
<tr>
<th>Generalize across</th>
<th>a</th>
<th>a,e</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>pa</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>pd</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>pda</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>pe</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>pea</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>pde</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>r</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>rxx</td>
<td>0.76</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Results for dimension level scores
Validity

- Historical focus in individual differences measurement is on delineation and measurement of situationally stable knowledge, skill, ability, and trait constructs
  - These are latent, unobservable individual attributes
  - “Psychological” constructs

- In ACs, we can be dealing with a different type of construct - “behavioral constructs”

- This has important implications for the layer of meaning we ascribe to AC validity results
Constructs in ACs

- “The focal constructs assessed in an assessment center...are defined as a constellation or group of behaviors that are specific, observable, and verifiable; that can be reliably and logically classified together; and that relate to job success. The term dimension is sometimes used synonymously with competency or KSA (knowledge, skills, or ability). Other assessment center applications have classified relevant behaviors according to tasks or job roles. Regardless of the label for the focal constructs to be assessed, they must be defined behaviorally, and as such are referred to hereafter as “behavioral constructs.”

International Taskforce on AC Guidelines, 2014
Psychologists have long acknowledged behavior is a function of the person, situation, and their interaction.

We often want to make inferences based on scores from *behaviorally-based* assessments, and in applied settings these inferences usually involve how those scores relate to how people will subsequently *behave* on the job.

If behavior is a function of the person, situation, and their interaction — not just situationally *stable* individual difference constructs — then wouldn’t assessment and evaluation approaches focused solely on the latter provide a deficient indicator of how people subsequently behave?
## ACs: Traditional vs. Interactionist Perspectives

<table>
<thead>
<tr>
<th>Traditional Individual Differences</th>
<th>Interactionist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercise-related variance seen as unreliable or contaminant-source of variance with regard to performance on a given dimension</td>
<td>Exercise-related variance seen as reliable and construct-relevant variance in dimension-relevant behavior exhibited in response to different job critical tasks/situations</td>
</tr>
<tr>
<td>Each exercise provides only trivial insight into the domain of behaviors that comprise a dimension</td>
<td>Each exercise provides important, unique insights into the domain of dimension-relevant behavior exhibited in response to different job critical tasks/situations</td>
</tr>
<tr>
<td>Analyzing WEDR when evaluating AC functioning is misleading. We should only analyze aggregated data.</td>
<td>Analyzing WEDRs is critical to evaluating AC functioning because it provides a comprehensive (unconfounded) decomposition of variance underlying AC scores – components can always be aggregated post-analysis</td>
</tr>
</tbody>
</table>
Implications for evaluating AC score validity

- Our approaches to defining and evaluating construct-relevant variance are dominated by the traditional individual differences paradigm
  - Understandable for traditional psychological constructs, but less acceptable for “behavioral constructs”?

- Construct-relevant variance may exist beyond “dimension factors”

- “Exercise factors” and “dimension-exercise interactions” do not necessarily represent construct-irrelevant variance

- **Key question:** Did you design your exercises to provide parallel measures of your dimensions?
Effects of Aggregation….

- Results framed from a traditional individual differences paradigm
  - Dimensions good
  - Exercises bad
  - “Error” bad

- Assume exercises were designed to provide parallel measures of dimensions (akin to items on a cognitive ability test)

- Comparable to paradigm underlying Sackett & Dreher (1982)

Dimension level scores reflect averaging 3 WEDRs across exercises
Effects of Aggregation...from a traditional individual differences paradigm
Assuming exercises were designed to provide parallel measures of dimensions

<table>
<thead>
<tr>
<th>Source</th>
<th>Level of Score</th>
<th>WEDR</th>
<th>Dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.g</td>
<td>0.17</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>d.s</td>
<td>0.05</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>e.g</td>
<td>0.03</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>e.s</td>
<td>0.31</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.44</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

% C-Relevant | 0.22 | 0.44 |
% C-Irrelevant | 0.34 | 0.27 |
% Error | 0.44 | 0.29 |

Effects of aggregation...from an interactionist paradigm
Assuming exercises were designed to provide shared and unique perspectives on dimensions

<table>
<thead>
<tr>
<th>Source</th>
<th>Level of Score</th>
<th>WEDR</th>
<th>Dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.g</td>
<td>0.17</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>d.s</td>
<td>0.05</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>e.g</td>
<td>0.03</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>e.s</td>
<td>0.31</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.44</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

% C-Relevant | 0.56 | 0.71 |
% C-Irrelevant | 0.0 | 0 |
% Error | 0.44 | 0.29 |

Dimension level scores reflect averaging 3 WEDRs across exercises
Summary and Key Takeaways
Revisiting our objectives

- **Objective 1**: Visualizing variance in AC scores
- **Objective 2**: Visualizing the effects of aggregation
- **Objective 3**: Discuss the layer of meaning we ascribe when tying variance partitioning results to issues of reliability and validity
- **Objective 4**: Discuss the practical implications of these observations for evaluating the quality of AC scores in research and practice
We often draw firewalls between validity and reliability, but these concepts are *layers of meaning* we ascribe to variance decomposition results and analytic techniques which are “blind” to that layer of meaning.

The two articles find similar things with regard to distribution of variance in AC scores and role of aggregation, but the *layer of meaning* attached to them differs.
Fundamental issue #2

- The *same* source of variance can legitimately be viewed as a source reliability or unreliability depending on…
  - One’s AC *measurement design*
  - *One’s desired generalizations*
  - One’s *use of AC scores*
  - The underlying distribution of variance and effect of aggregation doesn’t change, *but the layer of meaning* applied to it does
Fundamental issue #3

• The same source of variance can legitimately be viewed as a source of construct relevant or construct irrelevant variance depending on…
  • How one designed his/her AC exercises
    • Did you intend your exercises to provide parallel measures of a given dimension?
  • One’s measurement and evaluation paradigm
    • Traditional individual differences paradigm
    • Interactionist paradigm
  • The underlying distribution of variance and effect of aggregation doesn’t change, but the layer of meaning applied to it does
Key takeaways

- When interpreting past AC research that has modeled the latent structure of AC scores, realize that nearly ALL of it is based on data that have not been aggregated beyond the WEDR-level
  - Conclusions regarding AC score reliability and validity will likely change upon aggregation

- This does not mean we should stop analyzing WEDR data
  - Analyzing WEDR data allows us to obtain very detailed decompositions of variance – if we aggregate prior to analysis, we lose this information
  - We can always aggregate decomposition results post-analysis!
Key takeaways

- Don’t blindly interpret variance decomposition results – the same numbers can have VERY different meaning and implications for judging the quality of your AC scores

- Ask yourself: Did I intend my exercises to provide parallel measures of a given dimension?
  - This has fundamental implications for how you evaluate reliability and validity of the resulting scores
  - What is one person’s error is another person’s true score (and vice versa)
  - What is one person’s construct-irrelevant variance is another person’s construct-relevant variance (and vice versa)
Key takeaways

- Be wary of imprecise language…define what you mean
  - “Dimension-effects” and “exercise-effects”
    - Specific meaning of variance attributable to a given factor or variance component will depend on structure of model
    - Allowing dimension factors to correlate or constraining them to be uncorrelated fundamentally changes the meaning of those factors
  - “Reliable” vs. “unreliable” variance
    - Composition will depend on measurement design, desired generalization, and score use within a given study/application
  - “Construct–relevant” vs. “construct-relevant” variance
    - Composition will depend on how you designed your AC (how you intended exercises to function) and your measurement and evaluation paradigm
Key takeaways

- Consider the *International Taskforce on AC Guidelines, 2014:*

  “Although these Guidelines do not prescribe use of a specific type of score, as this will vary across assessment centers, what is paramount is that the validation evidence supporting the way in which the scores are ultimately used (in terms of their validity and reliability for the purpose at hand) is provided by the AC developer/user. **Whether these scores are exercise-specific dimension scores, across-exercise dimension scores, or some other type of aggregate score is not critical—what matters here is that the developer defends the validity of those scores in reference to how they are being used.**”

- Be wary of falling into the trap of developing and evaluating measures of “behavioral constructs” as if they were not behavioral in nature
Questions?


