

## Mysteries of I/O: What Are Overall Assessment Centre Ratings?

Presentation to the I/O SIG, 22<sup>nd</sup> February, 2007

Dr. Duncan Jackson  
Massey University  
Auckland

## Dr. Duncan Jackson

---

- Currently faculty member at Massey University, Albany, Auckland
- Teaches human resource management topics and general management
- Researches assessment centres (ACs) (was the focal topic for my Ph.D.)

## Assessment centres defined

---

- ACs involve the participation of [often] a group of candidates in multiple exercises, who are observed and rated by a team of trained assessors on predetermined task related behaviours and/or dimensions (Ballantyne & Povah, 1995)

## Assessment centres defined

---

- They often include supplementary psychometric tests in the hope of adding incremental validity
- E.g., personality, cognitive ability
  - I will focus on the simulation component

## A Measurement Issue

---

- Debate has raged about ACs since 1982
- Issues akin to the notion of adding apples and oranges

## In ACs

---

- Participants in ACs receive overall scores, based on trait-measures (called dimensions)

## ACs

---

- When factor analysed, we find that measures of the same dimension (e.g., tolerance) do not fit together well

## Simple Example

---

- Mechanic's position
- Three different exercises
- Three dimensions
  - Communication
  - Persistence
  - Tolerance

## Simple Example

---

### Group Exercise

Communication = 5  
Persistence = 5  
Tolerance = 5

### Work Sample

Communication = 1  
Persistence = 1  
Tolerance = 1

### Inspection Report

Communication = 3  
Persistence = 3  
Tolerance = 3

## Multitrait-Multimethod Problem

---

- We often find that correlations across the same dimension are moderate (monotrait-heteromethod)
- And correlations among different dimensions within exercises are strong (heterotrait-monomethod)

## Dimensions as Traits

---

- Note that the aim in traditional ACs is to measure dimensions as though they are stable traits
- That is, they are treated as characteristics that will be more or less stable across situations

## An AC from practice

---

- Entry-level customer service role
- AC used for employee selection
- 208 job applicants
- Team of trained assessors who were experienced on the focal job

## Steps in an AC Approach

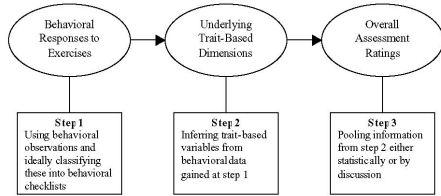


Figure 1. Assessment Steps, Chronologically, During an Assessment Center

## Example from a Real AC

- Three Exercises
  - Group discussion/problem solving format
1. Approaching customers (Ex1)
  2. Closing sales (Ex2)
  3. Returning goods (Ex3)

Note: Each exercise had an associated 10-item behavioural checklist. All variables were scored on a 1-6 scale.

## Example from a Real AC

- Five Dimensions
1. Comprehension (D1)
  2. Oral expression (D2)
  3. Tolerance (D3)
  4. Teamwork (D4)
  5. Customer focus (D5)

Table 1  
Multitrait-Multimethod Matrix for Assessment Dimensions and Exercises

	Ex1D1	Ex1D2	Ex1D3	Ex1D4	Ex1D5	Ex2D1	Ex2D2	Ex2D3	Ex2D4	Ex2D5	Ex3D1	Ex3D2	Ex3D3	Ex3D4
Ex1D1														
Ex1D2	<b>.64</b>													
Ex1D3	.49	<b>.65</b>												
Ex1D4	.57	.71	<b>.81</b>											
Ex1D5	.55	.62	.61	<b>.63</b>										
Ex2D1	.38	.42	.38	.38	.31									
Ex2D2	.32	.41	.38	.34	.26	<b>.62</b>								
Ex2D3	.17	.25	.33	.28	.18	.48	<b>.54</b>							
Ex2D4	.32	.38	.45	.43	.33	.55	.65	<b>.71</b>						
Ex2D5	.25	.30	.39	.37	.34	.56	.55	.53	<b>.71</b>					
Ex3D1	.20	.29	.24	.29	.26	.27	.35	.29	.38	<b>.42</b>				
Ex3D2	.30	.42	.34	.40	.34	.39	.45	.34	.43	.44	<b>.78</b>			
Ex3D3	.12	.24	.13	.21	.19	.23	.34	.34	.38	.41	.77	<b>.69</b>		
Ex3D4	.22	.34	.29	.35	.32	.35	.38	.33	.41	.45	.88	.81	<b>.79</b>	
Ex3D5	.20	.29	.23	.29	.28	.29	.36	.35	.45	.46	.84	.73	.80	<b>.84</b>

Note: D = dimension, D1 = comprehension, D2 = oral expression, D3 = tolerance, D4 = teamwork, D5 = customer focus. Ex = Exercise. All correlations were significant ( $p < .05$ ). Monotrait-heteromethod correlations appear in frames (median = .34, IQR = .13), heterotrait-heteromethod correlations appear in bold (median = .61, IQR = .09).

Which resulted in factor analyses (and confirmatory factor analyses) that look like ...

Table 2  
Factor Analysis of Assessment Center Ratings

Item	Factor			$h^2$	$M$	$SD$
	1	2	3			
Ex1D1	-.03	<b>.70</b>	.02	.48	4.99	0.79
Ex1D2	.05	<b>.82</b>	<-.01	.70	4.76	1.06
Ex1D3	-.07	<b>.77</b>	.13	.68	4.74	0.89
Ex1D4	.03	<b>.86</b>	.01	.77	4.57	1.02
Ex1D5	.07	<b>.76</b>	-.06	.56	5.05	0.84
Ex2D1	-.04	.16	<b>.63</b>	.50	4.93	0.90
Ex2D2	.02	.05	<b>.72</b>	.58	4.78	0.98
Ex2D3	-.01	-.11	<b>.81</b>	.57	4.82	0.88
Ex2D4	<.01	.02	<b>.88</b>	.79	4.63	0.95
Ex2D5	.13	.02	<b>.68</b>	.58	4.82	0.94
Ex3D1	<b>.95</b>	.02	-.06	.86	4.43	1.15
Ex3D2	<b>.76</b>	.15	.06	.73	4.41	1.18
Ex3D3	<b>.87</b>	-.12	.06	.74	4.45	1.02
Ex3D4	<b>.93</b>	.06	-.02	.90	4.12	1.17
Ex3D5	<b>.89</b>	-.02	.04	.82	4.50	1.11
SS	5.22	4.57	4.91			
%	44.69	59.71	68.37			

Note:  $h^2$  = Communality estimates upon extraction. D = dimension, D1 = comprehension, D2 = oral expression, D3 = tolerance, D4 = teamwork, D5 = customer focus. Ex = Exercise. SS = rotated sums of squared loadings. % = cumulative percent of variance explained. Factor correlations between 1 and 2 = .34, 1 and 3 = .50, and 2 and 3 = .50.

This is a well known effect: The *exercise effect*

---

- See Lance et al. (2004, JAP) and Jackson, et al. (2005, HP).
- The question remains: If ACs don't measure what they should, then why are they predictive?

## Note

---

- In personality theory, exercise effects are thought of as indicating:
  - Halo effects
  - Method-related error

## Note

---

- The traditional and prevailing view is that exercise effects are similarly indicative of error in ACs
  - Strikes me as slightly odd and mysterious, given, at their heart, they are *behavioural* measures ...

## Overall Assessment Ratings (OARs)

## Overall Assessment Ratings (OARs)

---

- Often it is OARs that are used to predict performance
- No known research has looked at the inside workings of OARs

---

	Exercise 1	Exercise 2	Exercise 3	Dimension Score
Comprehension				
Oral expression				
Tolerance				
Teamwork				
Customer focus				
<b>OAR</b>				

## The following steps occur in ACs

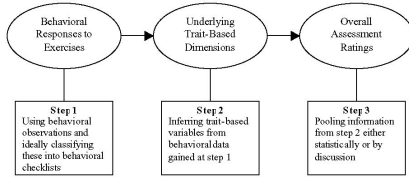


Figure 1. Assessment Steps, Chronologically, During an Assessment Center

## What do OARs reflect?

- We know about exercise effects
- Could OARs simply reflect behavioural performance?

## What is the relationship between Step 1 and Step 3?

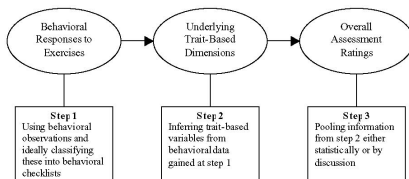


Figure 1. Assessment Steps, Chronologically, During an Assessment Center

- Or put another way, how much variation in dimension-based OARs can be explained by behavioural checklist scores?

- These OARs are *supposed* to be based on dimension assessments *across* exercises
- In this case, the OAR was based on an integration discussion

Behavioural score 1  
Behavioural score 2  
Behavioural score 3

Trait-based OARs

## MLR

- IVs = Average scores on three behavioural checklists
- DV = Dimension-based OAR
- Adjusted  $R^2 = .90$ 
  - $\beta(\text{Ex1}) = .37$
  - $\beta(\text{Ex2}) = .31$
  - $\beta(\text{Ex3}) = .51$

Note: All coefficients  $p < .001$

## HUH????? .90?????

---

- I mean, you would expect some shared variance, but 90%?
- I couldn't believe the relationship could be that strong!

## How about $r$ ?

---

- What is the simple correlation between:
  1. Average ratings across the three behavioural checklists ( $x$ ) and...
  2. Discussion-based OARs ( $y$ )

## And here it is...

---

- $r_{xy} = .95$  ( $p < .001$ )

## Repeats of the same effect

---

- I have since repeated this on other samples with similar results

## Implications for practice

---

- The findings of this research suggest the following:

## Implications for practice 1

---

- No matter how we try to transpose AC ratings into dimension categories, we cannot escape a behavioural assessment that is specific to exercises
  - Which implies ...

## Implications for practice 2

- Categorising behavioural information from ACs into dimension categories is possibly misleading and a waste of time

## This doesn't work:

	Exercise 1	Exercise 2	Exercise 3	Dimension Score
Comprehension				
Oral expression				
Tolerance				
Teamwork				
Customer focus				
<b>OAR</b>				

## But this might:

	Exercise 1	Exercise 2	Exercise 3
Behavioural item 1			
Behavioural item 2			
Behavioural item 3			
Behavioural item 4			
Behavioural item 5			
Behavioural item 6			
Behavioural item 7			
Behavioural item 8			
Behavioural item 9			
Behavioural item 10			
Exercise score			
Exercise OAR			

## On that note

- Worked with OPRA Consulting on a DC
- Correlation between scores on a behavioural DC and work performance ( $N = 106$ )
- Uncorrected correlation of .42
  - When corrected for attenuation due to unreliability and range restriction, the estimated correlation was .52
  - Meta-analyses of traditional ACs/DCs report corrected correlations of around .36 or .37

## Factor analysis

$N = 214$

Item	Factor					$R^2$	$M$	$SD$
	1	2	3	4	5			
Ex1b1				-.79		.69	3.33	1.29
Ex1b2				-.62		.67	3.74	1.28
Ex1b3				-.70		.55	3.37	1.10
Ex1b4				-.66		.62	3.29	1.23
Ex1b5				-.64		.48	3.34	1.23
Ex1b6				-.63		.40	3.01	1.08
Ex1b7				-.47		.34	3.89	1.08
Ex1b8				-.51		.64	3.32	1.35
Ex1b9				-.76		.71	3.34	1.21
Ex1b10				-.79		.74	3.09	1.29
Ex2b1		.81				.67	3.32	1.19
Ex2b2		.59				.68	3.29	1.41
Ex2b3		.58		.35		.55	3.89	1.04
Ex2b4		.57				.55	3.49	1.29
Ex2b5		.61				.37	3.45	1.12
Ex2b6		.81				.63	3.27	1.11
Ex2b7		.83				.74	3.23	1.24
Ex2b8		.70				.67	3.34	1.24
Ex2b9		.84				.72	3.35	1.21
Ex2b10		.67				.68	3.70	1.24
Ex3b1	.80					.66	2.91	1.27
Ex3b2	.72					.64	2.79	1.24
Ex3b3	.79					.64	2.88	1.20
Ex3b4	.77					.60	3.10	1.14
Ex3b5	.40					.23	3.69	1.07
Ex3b6	.75					.63	3.26	1.31
Ex3b7	.74					.64	3.15	1.44
Ex3b8	.82					.66	2.88	1.15
Ex3b9	.85					.78	2.77	1.20
Ex3b10	.79					.70	3.27	1.38
Ex4b1	.79					.71	3.10	1.32
Ex4b2	.74					.66	3.21	1.24
Ex4b3	.68					.49	3.43	1.32
Ex4b4	.86					.70	2.88	1.38
Ex4b5	.73					.62	2.86	1.36
Ex4b6	.75					.60	3.51	1.19
Ex4b7	.75					.58	3.22	1.31
Ex4b8	.80					.72	2.87	1.28
Ex4b9	.84					.72	2.97	1.22
Ex4b10	.72					.55	3.50	1.21

## Implications for practice 3

- That assessor discussions may not add value to the information obtained from ACs
  - This adds to a growing body of literature specific to ACs, but also in other areas of psychology, e.g.,
    - White, Michael J.; Spengler, Paul M.; Maugherman, Alan S.; Anderson, Linda A.; Cook, Robert S.; Nichols, Cassandra N.; Lampropoulos, Georgios K.; Walker, Blain S.; Cohen, Genna; Rush, Jeffrey D. (May, 2006) *The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction* Stefania Aegisdóttir. *Counseling Psychologist*, vol. 34, no. 3, pp. 341-382.
    - I think you guys need another author or two...

## Questions

---

- The MTMM problem implies that AC dimensions are treated like traits.
  - Should we focus on behavioural checklists and do away with the trait inferences?
  - Do we need to encourage trait-like assessment in ACs?
  - Are companies opening themselves up to legal challenges?
  - Does this suggest a wider problem about the use of competencies in HRM!!!?

## Are we doing the right thing?

---

- I do *not* believe that current practice, in this regard, is best practice
- I also believe that it's up to practitioners to take action

## Light Reading:

---

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, 18, 213-241.

Jackson, D. J. R., Barney, A. R., Stillman, J. A., & Kirkley, W. (in press). When traits are behaviors: The relationship between behavioral responses and trait-based overall assessment center ratings. *Human Performance*.

Lance, C. E. (in press). Why assessment centers don't work the way they're supposed to. *Interact/On*.

Stillman, J. & Jackson, D. J. R. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of Occupational and Organizational Psychology*, 78, 581-594.

Contact: [D.J.R.Jackson@massey.ac.nz](mailto:D.J.R.Jackson@massey.ac.nz)