**All analyses conducted in R. Code and output available upon request.**

1. *Below is an MTMM matrix representing data from a managerial assessment center (AC). Participants were rated on four dimensions (oral communication, team building, innovation, & stress tolerance) via three exercises (i.e., in-basket, allocation, & management problems). Please make the following calculations noted below. (20 points)*
2. *What is the degree of convergence (C)?*
3. *What is the degree of discrimination (D1 and D2)?*
4. *What is the degree of method variance (MV)?*
5. *Overall, does this AC seem to demonstrate appropriate construct-related validity?*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | Oral Communication | In-basket Exercise | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| 2 | Team Building | In-basket Exercise | 0.43 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| 3 | Innovation | In-basket Exercise | 0.38 | 0.48 | 1.00 |  |  |  |  |  |  |  |  |  |
| 4 | Stress Tolerance | In-basket Exercise | 0.31 | 0.22 | 0.30 | 1.00 |  |  |  |  |  |  |  |  |
| 5 | Oral Communication | Allocation Exercise | 0.63 | 0.40 | 0.39 | 0.27 | 1.00 |  |  |  |  |  |  |  |
| 6 | Team Building | Allocation Exercise | 0.37 | 0.55 | 0.42 | 0.22 | 0.51 | 1.00 |  |  |  |  |  |  |
| 7 | Innovation | Allocation Exercise | 0.30 | 0.40 | 0.53 | 0.17 | 0.45 | 0.51 | 1.00 |  |  |  |  |  |
| 8 | Stress Tolerance | Allocation Exercise | 0.32 | 0.20 | 0.24 | 0.72 | 0.38 | 0.28 | 0.24 | 1.00 |  |  |  |  |
| 9 | Oral Communication | Management Problems | 0.54 | 0.38 | 0.42 | 0.24 | 0.68 | 0.48 | 0.45 | 0.31 | 1.00 |  |  |  |
| 10 | Team Building | Management Problems | 0.36 | 0.44 | 0.44 | 0.22 | 0.49 | 0.62 | 0.41 | 0.28 | 0.60 | 1.00 |  |  |
| 11 | Innovation | Management Problems | 0.31 | 0.26 | 0.50 | 0.23 | 0.39 | 0.41 | 0.49 | 0.27 | 0.49 | 0.51 | 1.00 |  |
| 12 | Stress Tolerance | Management Problems | 0.34 | 0.21 | 0.28 | 0.70 | 0.34 | 0.28 | 0.22 | 0.80 | 0.38 | 0.32 | 0.30 | 1.00 |

This is a 4-trait, 4-method matrix. Ideally, SDDE > DDSE > DDDE

1. C1 = .6
2. D1 = .279  
   D2 = .206
3. MV = .074
4. Construct-related validity as analyzed by a MTMM matrix requires a judgement call. In this case, C1 indicates that 60% of the variance is explained by main effects of traits and shared variance specific to a given trait. MV indicates that variance due to method of measurement alone is 7.4%, which is much lower. For this reason, my evaluation is that **yes,** this assessment center demonstrates appropriate construct-related validity.
5. *The annotated output below is from a G-theory analysis of third grade students. The ratings are in relation to the aggressive behavior of these students. Each of the students (P) are rated by multiple teachers (F1) with multiple observations (F2). (25 points)*
   1. **Explain to me what each of the seven calculated facets are indicating (i.e., P, F1, F2, P\*F1, P\*F2, F1\*F2, & P\*F1\*F2).**

The ANOVA table conveys interaction effects between the subjects of the assessment and each facet involved. Of note is the “Proport” column, which expresses the proportion of variance in the ratings accounted for by each.

P – 26% of variance is explained by individual differences in people, outside the interaction of the individual with either facet 1 or facet 2. This portion is usable information yielded by the measurement process.

F1 – 9.9% of variance is explained by facet 1, outside the interaction of the individual with either individual differences or facet 2. Teachers may be rating differently than one another, which might also be measurable using inter-rater-reliability.

F2 – 2.4% of variance is explained by facet 2, outside the interaction of the individual with either individual differences or facet 1. Observation does not appear to be causing substantial systematic error.

P\*F1 – 18.7% of variance is explained by the interaction of individual differences with facet 1. Student-teacher relationships are likely to be causing systematic error. This is our largest measured source of error.

P\*F2 – 1.5% of variance is explained by the interaction of individual differences with facet 2. The interaction between student and observation does not appear to be causing substantial systematic error.

F1\*F2 - 8.6% of variance is explained by the interaction of facet 1 with facet 2. Teacher-observation interactions may be causing systematic error. Perhaps the teachers are going out to party Thursday night, causing problems with ratings in the Friday morning observation.

P\*F1\*F2 – 32.9% of variance is explained by the interaction of individual differences, facet 1, and facet 2. This is undifferentiated error and does not allow us to form additional conclusions.

1. **Overall, what do these results suggest regarding the functioning of this behavioral rating system?**

Currently, G is above .7 and Phi is below .7. In cases where rank ordering is important, but absolute scores or not, these are appropriate results and the system may be used as it stands. However, if absolute scores are important the administrators must increase instances of either Facet 1 or Facet 2 to raise the projected value of Phi.

1. **What changes could be made regarding the number of teachers rating to improve Phi to an acceptable level (.70)?**

Number of teachers could be increased to 6, raising Phi coefficient to .719.

1. **What changes could be made regarding the number of observations to improve Phi to an acceptable level (.70)?**

Number of observations could be increased to 4, raising Phi coefficient to .708.

1. **What is the best combination improvement to get Phi to an acceptable level?**

If observations are 1 hour long, here is the total cost in hours for the two options listed above:

Teachers 6, observations 2 = 12 person-hours

Teachers 4, Observations 4 = 16 person-hours

Increasing the number of teachers to 6 and fixing the number of observations at 2 is the most cost-efficient way to raise the projected Phi coefficient to an acceptable level. This is the best combination, as it yields adequate projected Phi while having considerably lower cost than any other option expressed in the D-study. Having more teachers participating also may reduce systematic error related to facet 1 indicated in the ANOVA table.

GENERALIZABILITY THEORY ANALYSES:

Design Type 3: two-facet fully-crossed design, as in P \* F1 \* F2

Number of persons/objects ('P'):

10

Number of levels for Facet 1 ('F1'):

4

Number of levels for Facet 2 ('F2'):

2

ANOVA Table:

df SS MS Variance Proport.

P 9.000 76.813 8.535 .781 .260

F1 3.000 31.938 10.646 .297 .099

F2 1.000 6.613 6.613 .071 .024

P\*F1 27.000 56.938 2.109 .561 .187

P\*F2 9.000 10.513 1.168 .045 .015

F1\*F2 3.000 10.738 3.579 .259 .086

P\*F1\*F2 27.000 26.638 .987 .987 .329

Error Variances:

Relative Absolute

.286 .429

G-coefficients:

G Phi

.732 .646

D-Study:

Entered D-Study values for Facet 1:

3 4 5 6 7

Entered D-Study values for Facet 2:

1 2 3 4 5

In the D-study results below, the levels of Facet 1 appear in

the first column, and the levels of Facet 2 appear in the first row.

D-Study G Coefficients

.000 1.000 2.000 3.000 4.000 5.000

3.000 .582 .676 .715 .736 .749

4.000 .644 .732 .767 .785 .797

5.000 .687 .770 .802 .819 .829

6.000 .720 .797 .827 .842 .852

7.000 .745 .818 .846 .860 .869

D-Study Phi Coefficients

.000 1.000 2.000 3.000 4.000 5.000

3.000 .488 .586 .627 .651 .665

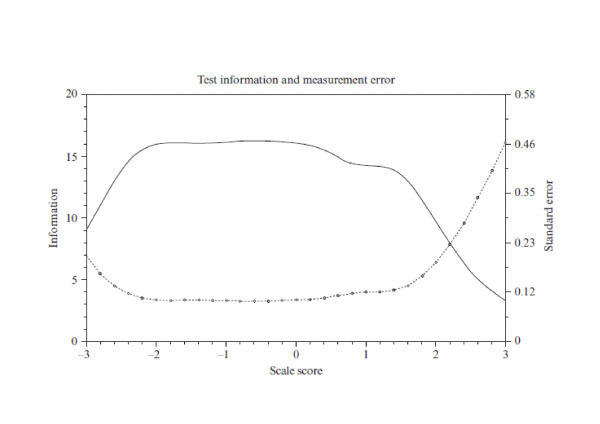
4.000 .548 .646 .686 .708 .722

5.000 .592 .688 .727 .748 .761

6.000 .625 .719 .757 .777 .790

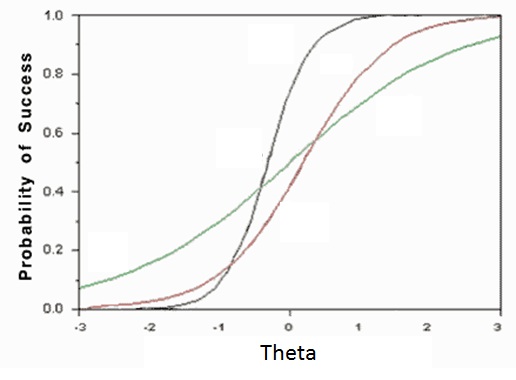
7.000 .652 .743 .779 .799 .811

1. *The figures below are from an item response analysis of a test of general mental ability (i.e., g or IQ). The first figure is the information curve for the entire test and the second figure includes the item characteristic curves for three items (i.e., A, B, & C). Looking at the first figure, what can you tell me about the information provided by this test? Compare and contrast the three different items. What can you tell me about each of those items? (25 points)*



**Figure one**

This test is positively skewed in measurement ability capturing lower and moderate levels of the latent trait General Mental Ability (*g*). This item information function figure shows a broad range, indicating that the scale is relatively low precision. I think it’s likely an information curve for a test of *g* in high-school students. This test can discriminate best between individuals ranging from -2 standard deviations below the mean to 1.5 standard deviations above the mean. Greater than -2 standard deviations below the mean, this scale exhibits more error and a reduction in ability to distinguish differences in *g.* Beyond 1.5 standard deviations above the mean, we see the scale’s ability to distinguish different levels of *g* decrease. Around 2.2 standard deviations above the mean, we see that the standard error of measurement is greater than the information – the ability to distinguish different levels of *g,* indicating that above that point the accuracy of the scale’s measurement is no better than chance.



B

C

A

**Figure two**

Item A

Location -.5 – This is low, indicating a respondent will be approximately ½ standard deviation below the mean on the trait measurement when getting the item correct 50% of the time.

Discrimination - Steep, indicating that this item is extremely related to the latent trait being measured.

Guessing – no guessing figure; very low or no chance of a respondent low on the parameter getting the item correct

Item B

Location – about .2 – This is low, indicating a respondent will be approximately 1/5 standard deviation above the mean on the trait measurement when getting the item correct 50% of the time.

Discrimination – Moderate, indicating that this item is moderately related to the latent trait being measured.

Guessing - no guessing figure; very low or no chance of a respondent low on the parameter getting the item correct

Item C

Location – about 0 – Indicating a respondent will be approximately average on the trait measurement and when getting the item correct 50% of the time.

Discrimination - Gradual, indicating that this item is only marginally related to the latent trait being measured.

Guessing - .08, indicating that respondents very low on the parameter may still get the answer correct.