Analyzing the structure of a test

Psychology 205

Department of Psychology
Northwestern University
Evanston, Illinois USA

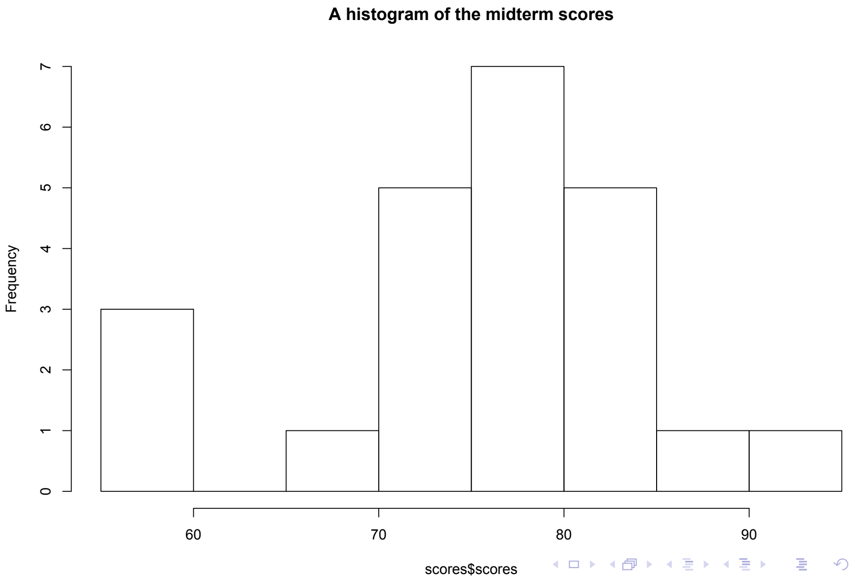
October, 2011

Outline

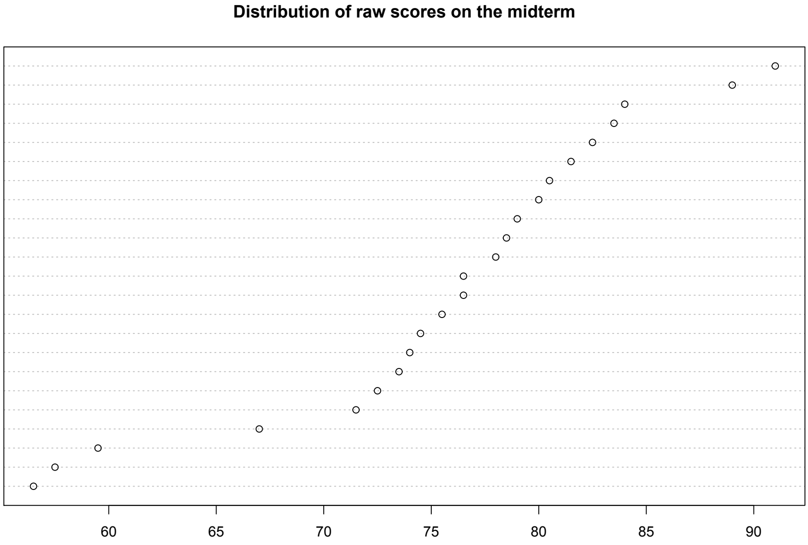
Analyzing a test: the example of the 205 midterm

- Code the data
  - For most tests, this will result in categorical scores
  - Usually with two categories, sometimes with 5-6 ordered categories
  - But for the midterm, the scores are ordinal
- Typical scores are just sums of the responses to each item
  - Sometimes these are weighted by item discrimination
- What is the internal consistency of the test?
  - Do the items all measure the same thing? (the latent construct)
  - What is amount of variance in the test associated with that central measure?
  - What would be the correlation of another test "just like it"
  - Multiple ways of estimating this, coefficients α & ω are two of the most common

The histogram of total scores



A dotchart of total scores



Does the test measure one central thing (construct)

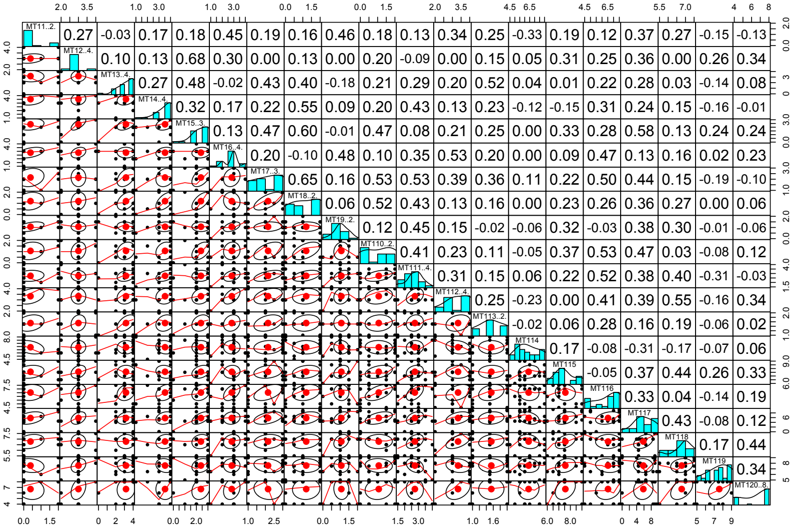
1. Reliability first discussed by Spearman (1904) as a limitation on test correlations
2. α Amount of test variance associated with average item
   - Cronbach (1951) and Guttman (1945) outlined one measure of a test reliability
3. Although α is the most common, other estimates are probably better.
   - α, β, and ω_h Revelle & Zinbarg (2009), Zinbarg, Revelle, Yovel & Li (2005),
4. All of these estimates are available in the *psych* package in R

First, describe the test items

```
> pairs.panels(mid)
> describe(mid)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
MT11..2.	1	23	0.39	0.78	0.0	0.26	0.00	0.0	2.0	2.0	1.43	0.90	0.16
MT12..4.	2	23	3.00	0.43	3.0	3.00	0.00	2.0	4.0	2.0	0.00	3.77	0.09
MT13..4.	3	23	3.17	1.03	3.0	3.32	1.48	0.0	4.0	4.0	-1.29	2.62	0.21
MT14..4.	4	23	3.52	0.79	4.0	3.68	0.00	1.0	4.0	3.0	-1.65	3.75	0.16
MT15..3.	5	23	2.39	0.78	3.0	2.53	0.00	0.0	3.0	3.0	-1.29	2.64	0.16
MT16..4.	6	23	2.83	0.72	3.0	2.84	0.00	1.0	4.0	3.0	-0.47	0.88	0.15
MT17..3.	7	23	2.15	0.82	2.0	2.18	1.48	1.0	3.0	2.0	-0.29	-1.42	0.17
MT18..2.	8	23	1.20	0.79	1.0	1.24	1.48	0.0	2.0	2.0	-0.23	-1.49	0.17
MT19..2.	9	23	1.07	0.43	1.0	1.08	0.00	0.0	2.0	2.0	-0.24	0.86	0.09
MT110..2.	10	23	1.07	0.79	1.5	1.08	0.74	0.0	2.0	2.0	-0.14	-1.64	0.16
MT111..4.	11	23	2.74	0.62	3.0	2.76	0.74	1.5	4.0	2.5	-0.23	0.16	0.13
MT112..4.	12	23	3.33	0.70	3.0	3.39	1.48	2.0	4.0	2.0	-0.54	-0.62	0.15
MT113..2.	13	23	1.54	0.37	1.5	1.55	0.74	1.0	2.0	1.0	-0.12	-1.01	0.08
MT114	14	23	6.39	1.05	6.0	6.39	0.74	4.5	8.0	3.5	0.22	-1.02	0.22
MT115	15	23	7.57	0.87	7.5	7.55	0.74	6.0	9.0	3.0	0.40	-0.57	0.18
MT116	16	23	6.52	1.08	7.0	6.63	0.74	4.5	7.5	3.0	-0.70	-0.91	0.23
MT117	17	23	5.83	2.66	5.5	6.08	1.48	0.0	9.5	9.5	-0.47	0.16	0.55
MT118	18	23	6.78	0.64	7.0	6.84	0.74	5.5	7.5	2.0	-0.83	0.13	0.13
MT119	19	23	7.59	1.25	8.0	7.71	1.48	5.0	9.0	4.0	-0.57	-0.36	0.26
MT120..8.	20	23	6.70	1.87	8.0	6.84	0.00	4.0	8.0	4.0	-0.69	-1.43	0.39

The interitem correlation matrix



Use the alpha function to score the test

```
alpha(mid,check.keys=FALSE)

Reliability analysis
Call: alpha(x = mid, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r mean   sd
      0.76      0.82    0.97      0.19  3.8 0.45

 Reliability if an item is dropped:
          raw_alpha std.alpha G6(smc) average_r
MT11..2.       0.75      0.82    0.96      0.19
MT12..4.       0.75      0.81    0.96      0.19
MT13..4.       0.75      0.82    0.95      0.19
MT14..4.       0.75      0.82    0.96      0.19
MT15..3.       0.73      0.80    0.96      0.17
MT16..4.       0.75      0.81    0.96      0.18
MT17..3.       0.74      0.80    0.96      0.18
MT18..2.       0.74      0.81    0.96      0.18
MT19..2.       0.76      0.82    0.96      0.19
MT110..2.      0.74      0.81    0.96      0.18
MT111..4.      0.74      0.81    0.95      0.18
MT112..4.      0.74      0.81    0.96      0.18
MT113..2.      0.75      0.82    0.96      0.19
MT114          0.78      0.84    0.97      0.21
MT115          0.74      0.81    0.96      0.19
MT116          0.74      0.81    0.96      0.18
MT117          0.75      0.80    0.96      0.18
MT118          0.74      0.81    0.95      0.18
MT119          0.78      0.83    0.97      0.21
MT120..8.      0.76      0.82    0.96      0.19
```

Item statistics for these 20 items

Item statistics

	n	r	r.cor	r.drop	mean	sd
MT11..2.	23	0.4291	0.422	0.2747	0.39	0.78
MT12..4.	23	0.4665	0.463	0.4508	3.00	0.43
MT13..4.	23	0.4392	0.434	0.3365	3.17	1.03
MT14..4.	23	0.4375	0.416	0.3092	3.52	0.79
MT15..3.	23	0.6998	0.705	0.6860	2.39	0.78
MT16..4.	23	0.5134	0.513	0.3716	2.83	0.72
MT17..3.	23	0.6561	0.655	0.5535	2.15	0.82
MT18..2.	23	0.6176	0.619	0.5333	1.20	0.79
MT19..2.	23	0.3782	0.370	0.2843	1.07	0.43
MT110..2.	23	0.5968	0.591	0.5482	1.07	0.79
MT111..4.	23	0.5994	0.605	0.4902	2.74	0.62
MT112..4.	23	0.5408	0.539	0.4753	3.33	0.70
MT113..2.	23	0.4473	0.430	0.3343	1.54	0.37
MT114	23	0.0065	-0.044	-0.1375	6.39	1.05
MT115	23	0.4964	0.482	0.4454	7.57	0.87
MT116	23	0.5680	0.564	0.4662	6.52	1.08
MT117	23	0.6699	0.669	0.5477	5.83	2.66
MT118	23	0.5201	0.524	0.4834	6.78	0.64
MT119	23	0.0773	0.048	-0.0096	7.59	1.25
MT120..8.	23	0.3756	0.368	0.2800	6.70	1.87

Use the alpha function–note that two items are actually negatively correlated with total

```
> alpha(mid)

Reliability analysis
Call: alpha(x = mid)
  raw_alpha std.alpha G6(smc) average_r mean   sd
      0.78      0.83    0.97       0.2  3.8 0.45

 Reliability if an item is dropped:
          raw_alpha std.alpha G6(smc) average_r
MT11..2.       0.77      0.82    0.96      0.20
MT12..4.       0.77      0.83    0.96      0.20
MT13..4.       0.77      0.83    0.95      0.20
MT14..4.       0.77      0.82    0.97      0.20
MT15..3.       0.76      0.82    0.96      0.19
MT16..4.       0.77      0.82    0.96      0.20
MT17..3.       0.76      0.81    0.96      0.19
MT18..2.       0.76      0.82    0.96      0.19
MT19..2.       0.77      0.83    0.96      0.20
MT110..2.      0.76      0.82    0.96      0.19
MT111..4.      0.76      0.82    0.95      0.19
MT112..4.      0.76      0.82    0.96      0.19
MT113..2.      0.77      0.83    0.96      0.20
MT114-         0.78      0.84    0.97      0.22
MT115          0.77      0.83    0.97      0.20
MT116          0.76      0.82    0.96      0.19
MT117          0.75      0.81    0.96      0.18
MT118          0.76      0.82    0.95      0.20
MT119-         0.79      0.84    0.97      0.22
MT120..8.      0.80      0.83    0.96      0.21
```

1. Consider what happens if an item is droped
2. Notice how the alpha goes up if 14 or 19 are dropped

A factor analysis can tell us which items correlate the most with the total score

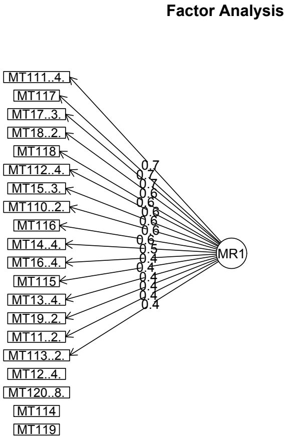
```
f1 <- fa(mid)
print(f1,sort=TRUE)

Factor Analysis using method =  minres
Call: fa(r = mid)
Standardized loadings based upon correlation matrix
           V   MR1     h2   u2
MT111..4. 11  0.75 0.5618 0.44
MT117     17  0.70 0.4938 0.51
MT17..3.   7  0.69 0.4750 0.52
MT18..2.   8  0.65 0.4177 0.58
MT118     18  0.61 0.3689 0.63
MT112..4. 12  0.60 0.3543 0.65
MT15..3.   5  0.58 0.3380 0.66
MT110..2. 10  0.57 0.3280 0.67
MT116     16  0.57 0.3239 0.68
MT14..4.   4  0.46 0.2117 0.79
MT16..4.   6  0.44 0.1952 0.80
MT115     15  0.42 0.1771 0.82
MT13..4.   3  0.42 0.1751 0.82
MT19..2.   9  0.40 0.1599 0.84
MT11..2.   1  0.39 0.1557 0.84
MT113..2. 13  0.36 0.1275 0.87
MT12..4.   2  0.29 0.0812 0.92
MT120..8. 20  0.27 0.0720 0.93
MT114     14 -0.09 0.0074 0.99
MT119     19 -0.07 0.0046 1.00

                MR1
SS loadings    5.03
```

1. MR1 is the correlation of the item with the factor
2. h^2 is the amount of variance explained in the item
3. u^2 is the amount of unexplained variance in the item

The factor analysis structural diagram



ω is an estimate of the general factor in a test

- Found by factoring the test
    - extracting 3 or more factors
- Then factoring these factors
    - Find a hierarchical solution
- How much does the general factor account for the test scores?

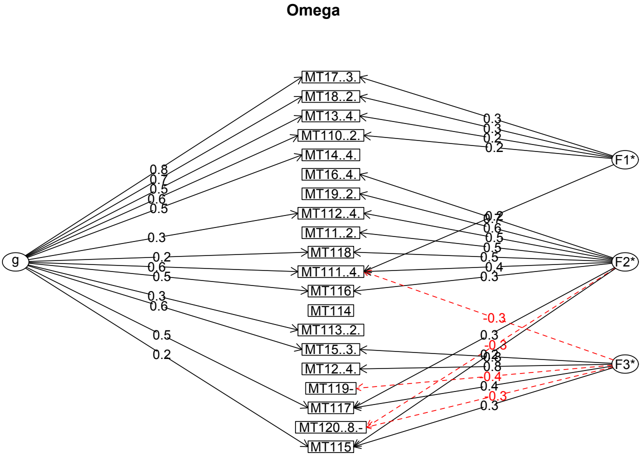
Omega output

```
Omega
Call: omega(m = mid)
Alpha:                 0.8
G.6:                   0.96
Omega Hierarchical:    0.59
Omega H asymptotic:    0.69
Omega Total            0.86

Schmid Leiman Factor loadings greater than  0.2
             g   F1*   F2*   F3*   h2   u2   p2
MT11..2.               0.51       0.31 0.69 0.09
MT12..4.                     0.80 0.68 0.32 0.01
MT13..4.    0.52  0.22            0.37 0.63 0.72
MT14..4.    0.47                  0.26 0.74 0.85
MT15..3.    0.56             0.81 1.00 0.00 0.31
MT16..4.               0.74       0.58 0.42 0.03
MT17..3.    0.76  0.30            0.68 0.32 0.85
MT18..2.    0.74  0.30            0.68 0.32 0.79
MT19..2.               0.58       0.37 0.63 0.06
MT110..2.   0.57  0.21            0.41 0.59 0.81
MT111..4.   0.61  0.24  0.42 -0.34 0.72 0.28 0.52
MT112..4.   0.33        0.53       0.39 0.61 0.28
MT113..2.   0.31                  0.13 0.87 0.73
MT114                             0.03 0.97 0.00
MT115       0.20        0.25  0.28 0.18 0.82 0.23
MT116       0.47        0.35       0.37 0.63 0.60
MT117       0.52        0.32  0.35 0.51 0.49 0.52
MT118       0.24        0.45       0.26 0.74 0.22
MT119-                       -0.45 0.25 0.75 0.14
MT120..8.-             -0.26 -0.34 0.19 0.81 0.00
```

1. α is the same as before
2. ω_h is the amount of general variance accounted for by the test

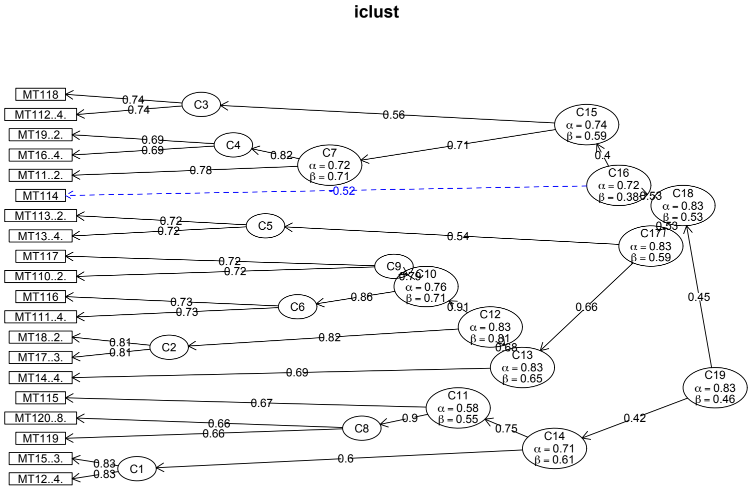
An omega diagram show three lower order factors and one general factor



Cluster analysis: Another way to group the items to show the structure

1. Form the item similarity matrix
   - This is just the interitem correlations
2. Combine the most similar pair
   - This is just simple addition
3. Repeat until some index does not increase

iclust of the 205 midterm



Why bother with all this analysis

1. A test score is only useful if it measures one central construct
   - We need to know how well the test measures that construct
   - By examining item statistics, we can try to improve the test
2. Classical test theory (as applied here) compares rank orders of people
   - If everyone knows the material to the same degree, then the test does not discriminate between people.
   - Internal consistency estimates of reliability are estimates of how much a test is measuring one central construct
3. Because tests are unreliable (not 100% reliable), we give multiple assessments.
4. Because tests are not perfectly valid (measuring the right construct), we give different types of assessments (e.g., papers, exams, in class contributions)
   - We need multiple measures and need to pool multiple sources of variance
5. Tests are not just assessments of knowledge, they are also sources of new knowledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*, *74*(1), 145–154.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H): Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.