What's Wrong with Factor-Analysing Tests Conforming to the Requirements of Item Response Theory?

John Raven,
30 Great King St,
Edinburgh EH3 6QH

Andy Fugard[1]
Department of Psychology,
University of Edinburgh

Version Date: 23 May 2008

ABSTRACT

Although many of those who are familiar with Item Response Theory (IRT) are well aware that factor analysing matrices of correlations between the items constituting such tests tends to yield misleading results, this is not as well known as it should be. In fact, endless researchers have come to seriously misleading conclusions as a result of applying factor analysis to such tests. The current paper illustrates just how misleading these results can be by factor analysing computer-generated data simulating that which would be obtained from the use of that ultimate form of an IRT test – a tape measure or meter stick – to measure height.

*****

The purpose of this paper is to illustrate, as dramatically as possible, something that is relatively well known to researchers familiar with applied Item Response Theory. This is that the application of the factor analytic procedures that are routinely used to establish the "unidimensionality" or otherwise of tests constructed according to Classic Test Theory yield "nonsense" when applied to tests conforming to Item Response Theory.

Such procedures *always,* and necessarily, declare that tests whose internal consistency, and, to some extent, unidimensionality, has been established using Item Response Theory are multi-dimensional. The procedure *always* points to the conclusion that three or more factors are required to account for the observed pattern of correlations between the items and that the test under investigation is therefore is multi-dimensional. The first part of this observation is correct. But it in no way supports the conclusion that the test is multi-dimensional. The "factors" that the procedure correctly indicates as being necessary to account for the maximum explainable variance in the correlation matrix are, in fact, "power" factors, each comprised of items of similar difficulty and split off from groups of easier or more difficult items. Failing to follow the recommendations of the APA Task Force on Statistical Inference (which would require researchers to look carefully at the *correlation matrix* which lay behind their factor analysis and ask what kind of model would fit it before undertaking their factor analyses) those concerned typically then examine the manifest content of the items with high "loadings" on each factor and label the factors on that basis ... although, in reality, the groups of items consist mainly of items of similar difficulty differentiated from "easier" or "more difficult" items.

[1] Andy would like to acknowledge the value of his discussion of these issues with Tim Bates.

The example we will use to illustrate the point will be a tape measure or meter stick used to measure length or height. This constitutes a perfect IRT scale. “People” “pass” (get right) all the centimetre marks up to that which registers their height and “fail” (are unable to reach) all the centimetre marks beyond that.

We will provide two illustrations of what happens.

One is based on computer-generated data approximating those that would have been obtained if a 36cm tape measure had been used to measure the heights of a species or strain of animals having a mean height of 18cm. That is, the computer was programmed to create a data set in which the mean would be 18 and the “scores” distributed across the entire 36 item scale according to a Gaussian (often misleadingly called a “normal”) distribution. Each “total score” ... ie “height” ... was calculated on the assumption that a particular animal would have “passed” each centimetre mark up to that point and failed to reach each centimetre mark beyond.

The second simulation specified a rectilinear distribution ... ie it specified that *the same number* of animals should get each “score” from 0 to 36. This would correspond to the distribution which might have been obtained had the tape measure been used to measure the heights of all animals and objects in a particular location ... or if a psychological test had been constructed to yield a Test Information Function curve which would reveal that the test had similar discriminative power over its entire operational range.

The first simulation yields data approximating those that are usually obtained by administering a typical IRT-based test to a cross section of respondents and factor analysing the results.

But the results of the second simulation illustrate the basic point to be made here even more clearly.

Table 1 shows the correlation matrix obtained by generating pass/fail data (“scores”) to fit a 36 centimetre tape measure with a mean of 18 and Gaussian distribution extending to the upper and lower limits of the tape measure.

Table 1

Correlations between pass/fail data for the centimetre marks on a 36 cm tape measure.

Computer generated data. Gaussian distribution of total "scores" with mean of 18. n = 1,000. Decimal point omitted.

"Respondents" "pass" all "items" up to their final "score" ("height") and "fail" all subsequent "items".

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	100																																			
2	32	100																																		
3	26	35	100																																	
4	31	40	51	100																																
5	26	34	49	62	100																															
6	23	26	43	53	61	100																														
7	25	31	43	55	62	69	100																													
8	22	30	39	51	57	63	72	100																												
9	18	21	34	45	48	56	63	70	100																											
10	21	24	33	43	47	54	63	67	73	100																										
11	21	22	31	39	44	49	55	62	66	73	100																									
12	20	19	25	35	40	44	52	55	63	69	73	100																								
13	19	20	26	37	44	45	53	57	63	67	72	79	100																							
14	15	17	22	30	36	38	49	52	57	59	63	69	78	100																						
15	14	17	22	29	35	35	46	50	54	59	61	66	74	77	100																					
16	14	16	20	27	32	35	43	44	50	54	56	62	70	70	77	100																				
17	12	17	19	26	31	33	40	44	47	50	50	58	65	66	73	76	100																			
18	14	16	18	26	25	31	36	37	42	45	48	52	58	59	68	70	74	100																		
19	13	14	16	21	25	30	34	36	40	43	47	52	57	60	67	68	71	75	100																	
20	12	13	15	21	25	29	33	35	39	41	44	51	55	58	64	64	68	70	77	100																
21	11	12	15	22	25	30	34	36	39	41	42	47	52	52	60	61	64	66	72	76	100															
22	8	12	16	19	20	25	26	28	33	34	39	43	48	49	55	57	60	61	67	72	74	100														
23	8	11	15	22	25	28	30	31	35	37	38	42	46	48	52	55	59	59	64	71	70	77	100													
24	10	9	17	20	23	23	25	29	32	34	35	40	43	45	50	50	55	55	62	66	66	71	76	100												
25	7	10	15	18	19	20	26	27	31	34	34	39	41	42	49	50	53	52	57	62	63	67	70	75	100											
26	5	8	11	17	16	20	21	23	28	29	30	34	37	36	43	45	49	48	53	57	56	63	65	68	75	100										
27	9	13	12	17	18	20	23	23	25	29	31	32	36	35	40	43	45	44	48	50	53	55	59	63	69	71	100									
28	9	9	9	12	12	15	18	18	22	25	28	31	32	32	37	39	41	39	45	49	49	52	52	58	66	67	70	100								
29	2	6	9	12	13	13	18	20	19	22	25	26	29	26	34	33	34	36	40	40	42	46	48	57	58	61	62	69	100							
30	4	3	7	11	10	14	13	15	16	17	21	21	21	23	27	26	27	28	33	32	35	37	37	44	47	50	50	57	60	100						
31	6	8	10	13	12	13	15	16	18	20	23	24	25	24	28	30	29	30	35	36	35	35	38	44	45	46	48	52	59	58	100					
32	-1	8	9	11	14	12	16	15	15	17	20	20	21	21	26	26	23	23	27	29	30	32	34	37	40	43	42	48	53	50	59	100				
33	-2	4	3	5	6	8	9	8	6	10	11	13	16	13	18	22	18	22	24	25	24	25	26	30	32	34	35	38	42	41	47	55	100			
34	7	5	4	7	8	8	10	11	8	10	10	9	12	10	12	14	12	11	14	15	15	16	18	19	25	25	22	28	31	29	32	36	47	100		
35	-2	-2	3	0	3	2	4	3	5	7	5	10	9	8	9	11	7	11	12	12	9	10	9	11	13	16	14	11	17	20	16	21	32	29	100	
36	3	7	-2	2	0	-1	2	1	-3	-7	-4	-5	-3	1	-2	-3	-2	0	-2	0	-1	1	1	0	1	-3	1	-1	-4	-4	-2	-1	-7	-5	0	100

The correlations adjacent to the diagonal tend to .99 because it is a fairly safe bet that if an animal is, for example, more than 1 cm tall it will also be more than 2 cm tall. So one can predict well from one to the other: the correlation is high. On the other hand they tend to zero in the distal corner of the matrix because knowing whether an animal is more than 1cm tall tells one very little about whether it will be more or less than 36 cm tall. (The deviations from such an "ideal" matrix will be explained later).

Fitting a "single factor" factor-analytic solution to these data yields the results below:

Table 2

Factor Loadings obtained by calling up a Single-Factor solution when applying Factor Analysis to the correlation matrix in Table 1.

Final Score/ cm. mark	Loadings on Factor 1
1	0.19
2	0.23
3	0.29
4	0.38
5	0.43
6	0.47
7	0.54
8	0.57
9	0.62
10	0.65
11	0.67
12	0.72
13	0.77
14	0.76
15	0.81
16	0.81
17	0.81
18	0.79
19	0.81
20	0.81
21	0.79
22	0.76
23	0.76
24	0.74
25	0.72
26	0.67
27	0.63
28	0.59
29	0.53
30	0.44
31	0.46
32	0.40
33	0.31
34	0.22
35	0.14
36	
SS Loadings	13.24
Proportion Var	0.37

Most researchers steeped in factor analysis but not familiar with Item Response Theory would interpret these results to mean, not merely that the correlation matrix cannot be "explained" by a single underlying factor (which would have been obvious if they had followed the recommendations of the APA Task Force on Statistical Inference and examined their correlation matrix before subjecting it to factor analysis), but also that the test is not "unidimensional".

They would then proceed to extract more factors.

The results of a 3 factor solution are shown in Table 3.

Table 3

3-factor solution from factor analysing the correlation matrix in Table 1.

Final Score/		Loadings on:	
cm. mark	Factor 1	Factor 2	Factor 3
1		0.30	
2		0.37	
3		0.51	
4		0.64	0.12
5	0.10	0.69	
6	0.14	0.71	0.10
7	0.20	0.79	
8	0.25	0.78	
9	0.35	0.72	
10	0.39	0.72	
11	0.45	0.65	
12	0.56	0.58	
13	0.63	0.57	
14	0.68	0.46	
15	0.74	0.40	0.15
16	0.75	0.33	0.18
17	0.77	0.28	0.21
18	0.76	0.21	0.23
19	0.77	0.17	0.31
20	0.76	0.14	0.37
21	0.69	0.15	0.41
22	0.66		0.48
23	0.61	0.12	0.53
24	0.54	0.10	0.62
25	0.48	0.10	0.68
26	0.41		0.72
27	0.33	0.12	0.72
28	0.28		0.75
29	0.19	0.10	0.76
30	0.12		0.66
31	0.14	0.12	0.64
32		0.13	0.61
33			0.51
34			0.39
35			0.21
36			
SS Loadings	7.47	6.01	5.85
Proportion Var	0.21	0.17	0.16
Cumulative Var	0.21	0.37	0.54

A glance back at Table 1 shows what has happened.

Speaking loosely, and skipping over the technicalities which distinguish factor analysis from cluster analysis, what a multiple-factor factor analysis “tries“ to do is to form groups of items that have high correlations with other items in the same cluster and low correlations with the items in other clusters.

So, what the program has, in effect, done is say “Look. Here in the middle there is a bunch of items that correlate highly with each other and relatively less with the two other bunches of items at the bottom and top ends of the scale. So, guys, you need at least 3 factors to account for these data”.

The way in which the program has “grouped” the items is shown in Table 4.

Table 4

Correlations in Table 1 Grouped into Clusters as indicated by 3-Factor analysis.
Decimal point omitted.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	100																																			
2	32	100																																		
3	26	35	100																																	
4	31	40	51	100																																
5	26	34	49	62	100																															
6	23	26	43	53	61	100																														
7	25	31	43	55	62	69	100																													
8	22	30	39	51	57	63	72	100																												
9	18	21	34	45	48	56	63	70	100																											
10	21	24	33	43	47	54	63	67	73	100																										
11	21	22	31	39	44	49	55	62	66	73	100																									
12	20	19	25	35	40	44	52	55	63	69	73	100																								
13	19	20	26	37	44	45	53	57	63	67	72	79	100																							
14	15	17	22	30	36	38	49	52	57	59	63	69	78	100																						
15	14	17	22	29	35	35	46	50	54	59	61	66	74	77	100																					
16	14	16	20	27	32	35	43	44	50	54	56	62	70	70	77	100																				
17	12	17	19	26	31	33	40	44	47	50	50	58	65	66	73	76	100																			
18	14	16	18	26	25	31	36	37	42	45	48	52	58	59	68	70	74	100																		
19	13	14	16	21	25	30	34	36	40	43	47	52	57	60	67	68	71	75	100																	
20	12	13	15	21	25	29	33	35	39	41	44	51	55	58	64	64	68	70	77	100																
21	11	12	15	22	25	30	34	36	39	41	42	47	52	52	60	61	64	66	72	76	100															
22	8	12	16	19	20	25	26	28	33	34	39	43	48	49	55	57	60	61	67	72	74	100														
23	8	11	15	22	25	28	30	31	35	37	38	42	46	48	52	55	59	59	64	71	70	77	100													
24	10	9	17	20	23	23	25	29	32	34	35	40	43	45	50	50	55	55	62	66	66	71	76	100												
25	7	10	15	18	19	20	26	27	31	34	34	39	41	42	49	50	53	52	57	62	63	67	70	75	100											
26	5	8	11	17	16	20	21	23	28	29	30	34	37	36	43	45	49	48	53	57	56	63	65	68	75	100										
27	9	13	12	17	18	20	23	23	25	29	31	32	36	35	40	43	45	44	48	50	53	55	59	63	69	71	100									
28	9	9	9	12	12	15	18	18	22	25	28	31	32	32	37	39	41	39	45	49	49	52	52	58	66	67	70	100								
29	2	6	9	12	13	13	18	20	19	22	25	26	29	26	34	33	34	36	40	40	42	46	48	57	58	61	62	69	100							
30	4	3	7	11	10	14	13	15	16	17	21	21	21	23	27	26	27	28	33	32	35	37	37	44	47	50	50	57	60	100						
31	6	8	10	13	12	13	15	16	18	20	23	24	25	24	28	30	29	30	35	36	35	35	38	44	45	46	48	52	59	58	100					
32	-1	8	9	11	14	12	16	15	15	17	20	20	21	21	26	26	23	23	27	29	30	32	34	37	40	43	42	48	53	50	59	100				
33	-2	4	3	5	6	8	9	8	6	10	11	13	16	13	18	22	18	22	24	25	24	25	26	30	32	34	35	38	42	41	47	55	100			
34	7	5	4	7	8	8	10	11	8	10	10	9	12	10	12	14	12	11	14	15	15	16	18	19	25	25	22	28	31	29	32	36	47	100		
35	-2	-2	3	0	3	2	4	3	5	7	5	10	9	8	9	11	7	11	12	12	9	10	9	11	13	16	14	11	17	20	16	21	32	29	100	
36	3	7	-2	2	0	-1	2	1	-3	-7	-4	-5	-3	1	-2	-3	-2	0	-2	0	-1	1	1	0	1	-3	1	-1	-4	-4	-2	-1	-7	-5	0	100

Of course, we could go on to, and, indeed *did* go on to, extract 5 factors.

But we have done enough to make the point: We *know* that the tape measure is unidimensional. Applying procedures derived from classic test theory to data obtained from it in an effort to establish whether or not it is unidimensional misleads. If pressed, factor analysis groups together items of similar *difficulty* and declares that they represent underlying factors or "dimensions" within the test. From the days of Guttman (who is best known for his work on "Scaleogram" analysis which is, in fact, a variant of IRT) onward these factors have been known as "power" factors.

But, failing to notice this, thousands of researchers who were not familiar with the objectives and measurement model lying behind IRT-based tests, and who failed to follow the recommendations of the APA task force on Statistical Inference (which, admittedly, may not have been published when they did their research) to "first look at your data", have then proceeded to commit a heinous crime which has influenced the thinking of generations of researchers.

As previously indicated, what they did was examine the manifest content of the items that had high loadings on these 3 or 5 factors and, from this, conclude that there were 3 or 5 (or more) "types" of item in the test ... in other words, the test was a mess and conflated 3, 5, or more "independent" dimensions[1, 2, 3].

A re-run

The above story corresponds in all essential details with what actually happens when researchers apply factor analysis to the matrices of correlations obtained by inter-correlating the items of IRT-based tests, so most readers will already have gained everything they may usefully learn from this article.

However, there is something puzzling about the correlation matrix shown in Table 1: Why are the correlations adjacent to the diagonal so far from .99 at the upper and lower ends of the scale?

The answer is that, because the simulation data have been generated to yield a Gaussian distribution over the length of the tape measure, there are few "respondents" to the "easiest" and "most difficult" "items".

So the analyses were re-run with a sample yielding the *same number* of "animals" having each "height" from 1 to 36 cm.

The results are shown in Tables 5 and 6.

Table 5

Correlations between pass/fail data for the centimetre marks on a 36 cm tape measure.

Computer generated data *with equal numbers* reaching each mark (and no further) from 1 to 36. n=3,700

("Respondents" "pass" all "items" [centimetre marks] up to that which indicates their height and "fail" all subsequent "items".) Decimal point omitted.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	100																																			
2	70	100																																		
3	56	80	100																																	
4	48	69	85	100																																
5	42	60	75	88	100																															
6	38	54	68	79	90	100																														
7	35	49	61	72	82	91	100																													
8	32	46	57	66	75	84	92	100																												
9	29	42	52	61	70	78	85	93	100																											
10	27	39	49	57	65	72	79	86	93	100																										
11	26	37	46	54	61	68	74	81	87	94	100																									
12	24	35	43	50	57	64	70	76	82	88	94	100																								
13	23	32	40	47	54	60	66	71	77	83	88	94	100																							
14	21	31	38	45	51	56	62	67	73	78	83	89	94	100																						
15	20	29	36	42	48	53	58	64	69	74	79	84	89	94	100																					
16	19	27	34	40	45	50	55	60	65	70	75	79	84	89	95	100																				
17	18	26	32	38	43	48	52	57	61	66	71	75	80	85	90	95	100																			
18	17	25	31	36	41	45	50	54	58	63	67	71	76	80	85	90	95	100																		
19	16	23	29	34	38	43	47	51	55	59	63	67	72	76	80	85	90	95	100																	
20	15	22	27	32	36	41	45	48	52	56	60	64	68	72	76	80	85	90	95	100																
21	15	21	26	30	35	38	42	46	49	53	57	60	64	68	72	76	80	85	90	95	100															
22	14	20	25	29	33	36	40	43	47	50	54	57	61	64	68	72	76	80	85	90	95	100														
23	13	19	23	27	31	34	38	41	44	47	51	54	57	61	64	68	72	76	80	85	89	94	100													
24	12	18	22	26	29	32	36	39	42	45	48	51	54	57	61	64	68	72	76	80	84	89	94	100												
25	12	17	21	24	27	30	33	36	39	42	45	48	51	54	57	60	64	67	71	75	79	84	89	94	100											
26	11	16	19	23	26	29	31	34	37	40	42	45	48	51	54	57	60	63	67	71	75	79	83	88	94	100										
27	10	15	18	21	24	27	29	32	35	37	40	42	45	47	50	53	56	59	63	66	70	74	78	83	88	94	100									
28	9	14	17	20	22	25	27	30	32	35	37	39	42	44	47	49	52	55	58	61	65	69	73	77	82	87	93	100								
29	9	13	16	18	21	23	25	28	30	32	34	36	39	41	43	46	48	51	54	57	60	64	67	71	76	81	86	93	100							
30	8	12	14	17	19	21	23	25	27	29	31	33	36	38	40	42	45	47	50	52	55	58	62	66	70	74	79	85	92	100						
31	7	11	13	15	17	19	21	23	25	27	29	30	32	34	36	38	41	43	45	48	50	53	56	60	64	68	72	78	84	91	100					
32	7	9	12	14	16	17	19	21	22	24	26	27	29	31	33	35	36	38	41	43	45	48	51	54	57	61	65	70	75	82	90	100				
33	6	8	10	12	14	15	17	18	20	21	23	24	26	27	29	30	32	34	36	38	40	42	45	47	50	54	57	61	66	72	79	88	100			
34	5	7	9	10	12	13	14	16	17	18	19	21	22	23	25	26	27	29	31	32	34	36	38	40	43	46	49	52	57	61	68	75	85	100		
35	4	6	7	8	9	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	29	31	32	35	37	39	42	46	49	54	60	69	80	100	
36	3	4	5	6	7	7	8	9	9	10	11	12	12	13	14	15	15	16	17	18	19	20	21	23	24	26	27	29	32	35	38	42	48	56	70	100

Table 6

5-factor solution from factor analysis of data in Table 5.

Final Score/ cm. mark	Loadings on: Factor 1	Factor 2	Factor 3	Factor 4
1				0.46
2				0.64
3		0.13		0.77
4		0.19		0.85
5	0.11	0.25		0.89
6	0.12	0.34		0.87
7	0.13	0.44		0.81
8	0.15	0.54		0.72
9	0.16	0.64		0.63
10	0.18	0.73		0.53
11	0.20	0.80		0.44
12	0.23	0.85	0.11	0.36
13	0.27	0.86	0.13	0.30
14	0.31	0.85	0.15	0.26
15	0.37	0.81	0.16	0.23
16	0.43	0.76	0.17	0.21
17	0.49	0.70	0.18	0.20
18	0.56	0.64	0.19	0.20
19	0.64	0.56	0.20	0.19
20	0.70	0.49	0.20	0.18
21	0.76	0.43	0.21	0.17
22	0.81	0.37	0.23	0.16
23	0.85	0.31	0.26	0.15
24	0.86	0.27	0.30	0.13
25	0.85	0.23	0.36	0.11
26	0.80	0.20	0.44	
27	0.73	0.18	0.53	
28	0.64	0.16	0.63	
29	0.54	0.15	0.72	
30	0.44	0.13	0.81	
31	0.34	0.12	0.87	
32	0.25	0.11	0.89	
33	0.19		0.85	
34	0.13		0.77	
35			0.64	
36			0.46	

The story is similar to that we obtained earlier except that the correlations between the “easiest” and “most difficult” “items” are much higher.

It may be thought that these results have no relevance to the main points being made in this article.

But that would not be true.

It is, in fact, extremely common for researchers to run item analyses, whether based on Classic test theory or Item Response Theory, using data derived from testing populations (often misleadingly called “samples”) which yield only a very narrow range of scores. This means that there are too few people – often no-one at all! – with scores in the tails of the distribution to permit the calculation of meaningful item statistics. This greatly exacerbates the errors made when the researchers concerned set about trying to interpret their results ... especially their factor analyses.

Notes

1. Actually, the use of the word "independent" itself reveals more than a little ignorance of how factor analysis works, because successive factors are not, in reality, independent of those which have gone before but actually represent the next best attempt to correct, with another single factor, the errors that have been made by assuming that the correlation [actually co-variance] matrix can be "explained" in terms of the previously extracted factors.

2. The term "unidimensional" is itself extremely ambiguous. For a discussion see eg Hattie, J. (1985) Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement* 9 139-164.

3. This error has led endless researchers to conclude that that the *Raven Progressive Matrices* measures several different things. While there may be senses in which this may be true, the point here is that this is not demonstrated by factoring the items. This is not the place to embark on a detailed discussion of these issues but, since the authors were led to prepare this paper by their interest in them, it is worth remarking that a review of the numerous IRT-based demonstrations that the various qualitatively different types of item which constitute the RPM measure the *same* underlying continuum of ability (in the sense that the qualitatively different types of item constituting the geological scale used to measure "hardness" all measure the same underlying variable) will be found in Raven J. & Raven, C.J. (2008) *Uses and Abuses of Intelligence.*(New York: Royal Fireworks Press) and that an example of the kind of study that might reveal other "dimensions" of difficulty (analogous to additional dimensions in the hardness of bricks) might be DeShon, R.P., Chan, D., & Weissbein, D.A. (1995). Verbal Overshadowing Effects on Raven's Advanced Progressive Matrices: Evidence for Multidimensional Performance Determinants. *Intelligence*, 21, 135-155.