
An introduction to principal component analysis and data reduction methods

**Vanderbilt-Ingram Cancer Center Biostatistics Workshop
October 17, 2008**

**Irene D. Feurer, Ph.D.
Research Professor of Surgery and Biostatistics
Vanderbilt University Medical Center
irene.feurer@vanderbilt.edu**

Workshop objectives

Present the basics of latent variable modeling as they relate to

- **survey instrument or test scaling and validation**
- **data reduction**
- **structural equation modeling**

Take the principal component model as a starting point and contrast it to a multiple group method

- **bi-factor analysis**
- **multiple group factor analysis**
- **cluster analysis**

Examples

■ Survey instrument scaling

- VTCPSI, Vanderbilt Transplant Center Patient Satisfaction Inventory
- QDQ, Vanderbilt Bill Wilkerson Center for Otolaryngology Quantitative Dizziness Questionnaire (*GP Jacobson, et al., unpublished data*)
- CES-D, Center for Epidemiological Studies Depression Scale

■ Data reduction

- Composite outcome measure (Pinson et al. *Ann Surg*, 2000;232:597-607.)
- Microarray data analysis (Wang & Gehan. *Statis Med.* 2005;24:2069-2087. Chen, Wang, Smith, Zhang. *Bioinformatics*. E-pub 2008 Aug 27.)

Latent variable methods

		Manifest Variables	
		Continuous	Categorical
Latent Variables	Scale of measurement		
	Continuous	factor analysis	latent trait analysis factor analysis of categorical data
Categorical	latent profile analysis analysis of mixtures	latent class analysis	

Basic terminology

- **Latent and manifest variables**
- **Exploratory vs. confirmatory factor analysis**
- **Data matrix**
- **Association matrix**
- **Principal component and common factor models**
- **Common and unique variance**
- **Scree plot of eigenvalues**
- **Factor loading**
- **Factor saturation**
- **Factor rotation: orthogonal and oblique**

Some historical notes

C Spearman. General intelligence, objectively determined and measured. *Am J Psychol* (1904).

C Spearman, C Burt, K Pearson, G Thompson, JCM Garnett, K Holzinger, early 1920's, mathematical foundations.

LL Thurstone, U Chicago, simple structure, mid 1930's.

Holzinger & Swineford, “bi-factor” model, *Psychometrika* (1937).

H Hotelling, principal components model, *J Educ Psychol* (1933).

S Mulaik, H Harman, R Cattell (scree plot), D Lawley, A Maxwell, A Baggaley, 1970's.

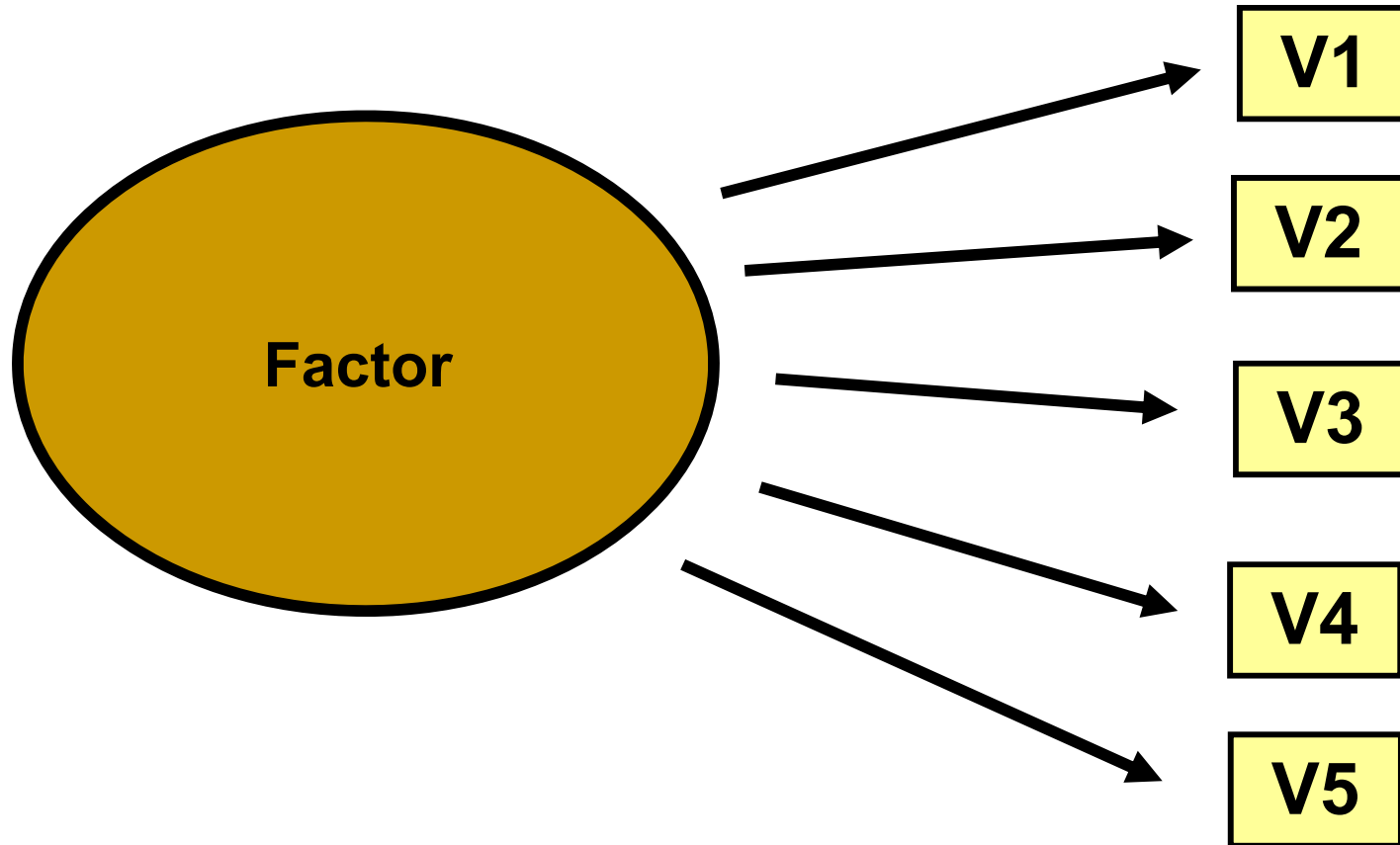
W Velicer, J Fava, R Gorsuch; simulation studies

R Gibbons & D Hedeker, dichotomous bi-factor model with IRT applications, *Psychometrika* (1992; 57:423-426).

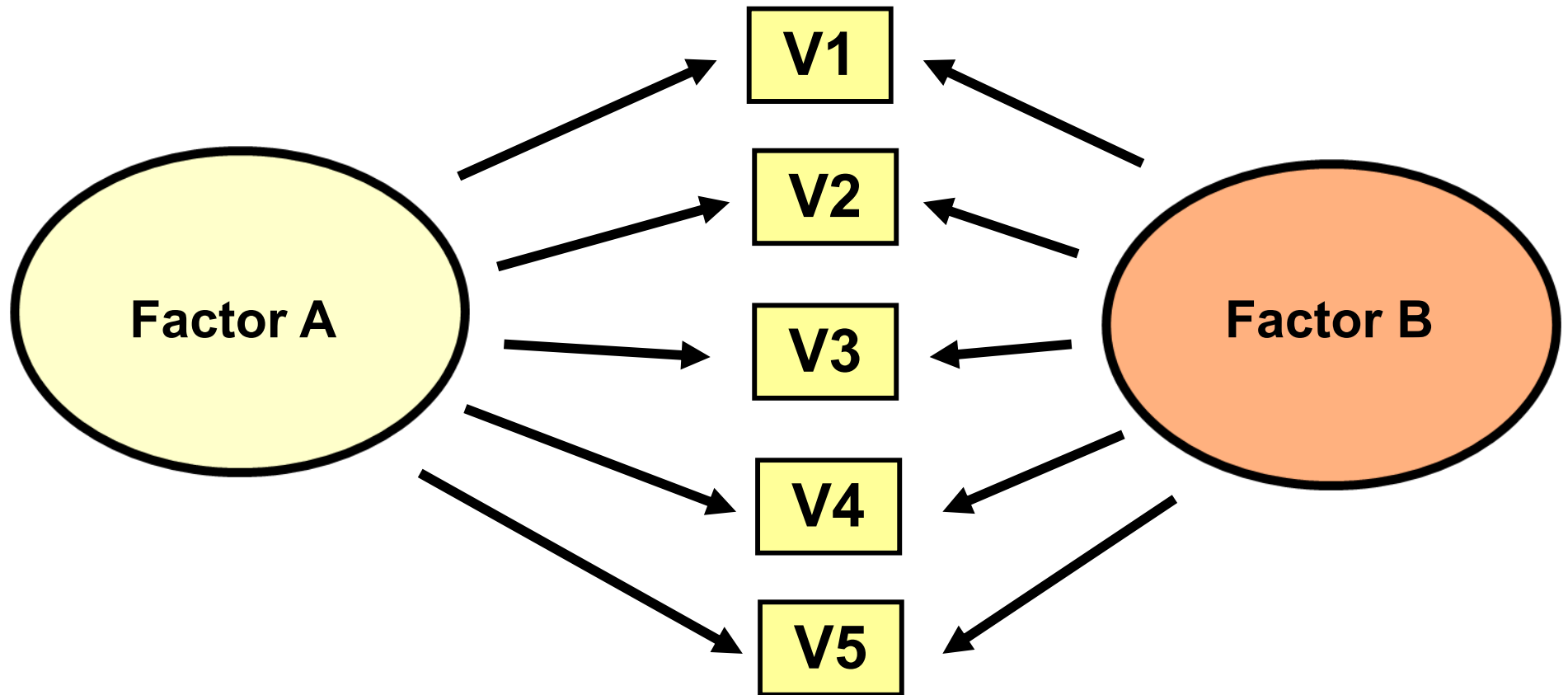
P Bentler (EQS), MW Browne, K Jöreskog (LISREL), R McDonald, N Waller (MicroFACT), R Gibbons, S Schilling, E Muraki, D Bock (TESTFACT); analysis of covariance structures, structural equation modeling, IRT applications; 1990's – .

Wang A, Gehan EA, Chen X, Wang L, Smith JD, Zhang B and others. Microarray data analysis; 2000's - .

Single latent variable (factor/component) with five manifest variables

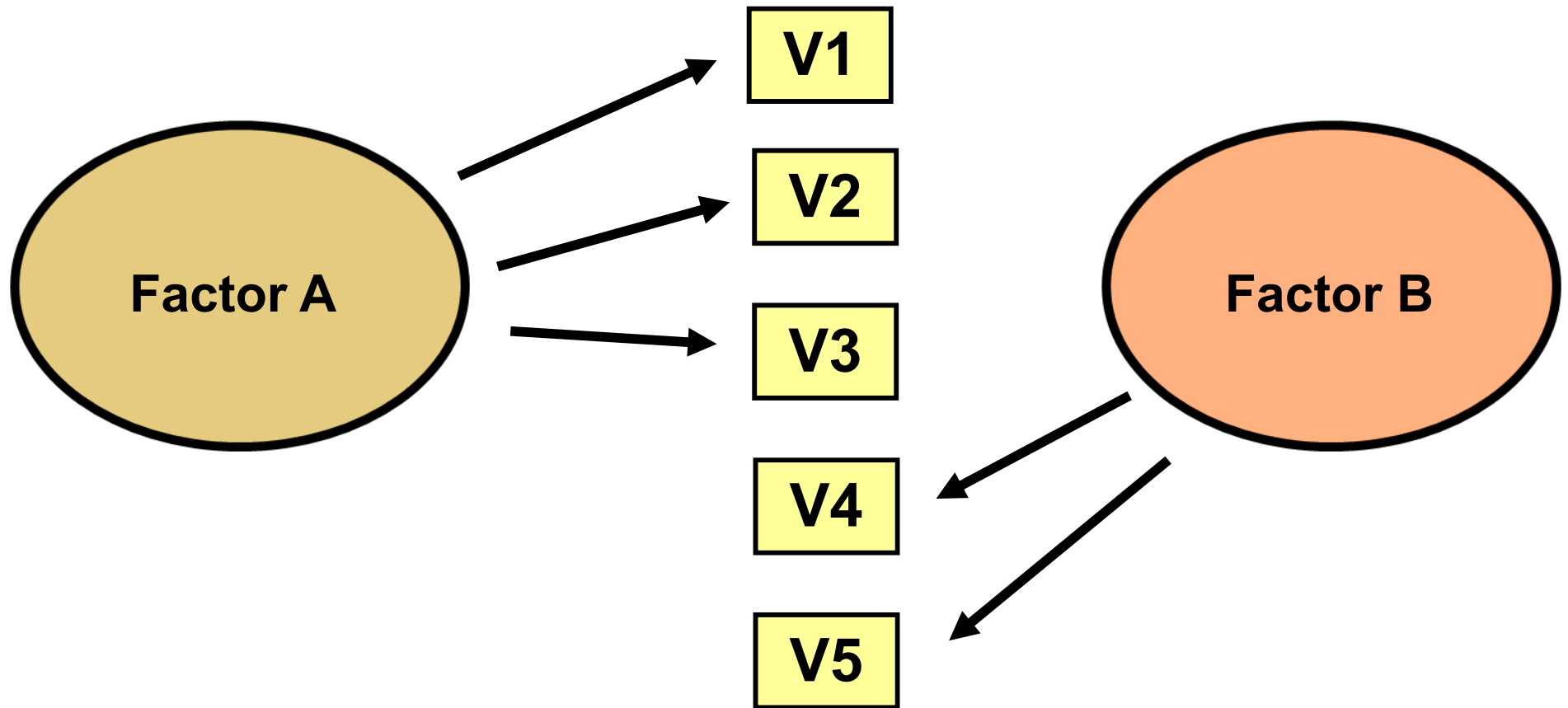


Two latent variables and five manifest variables

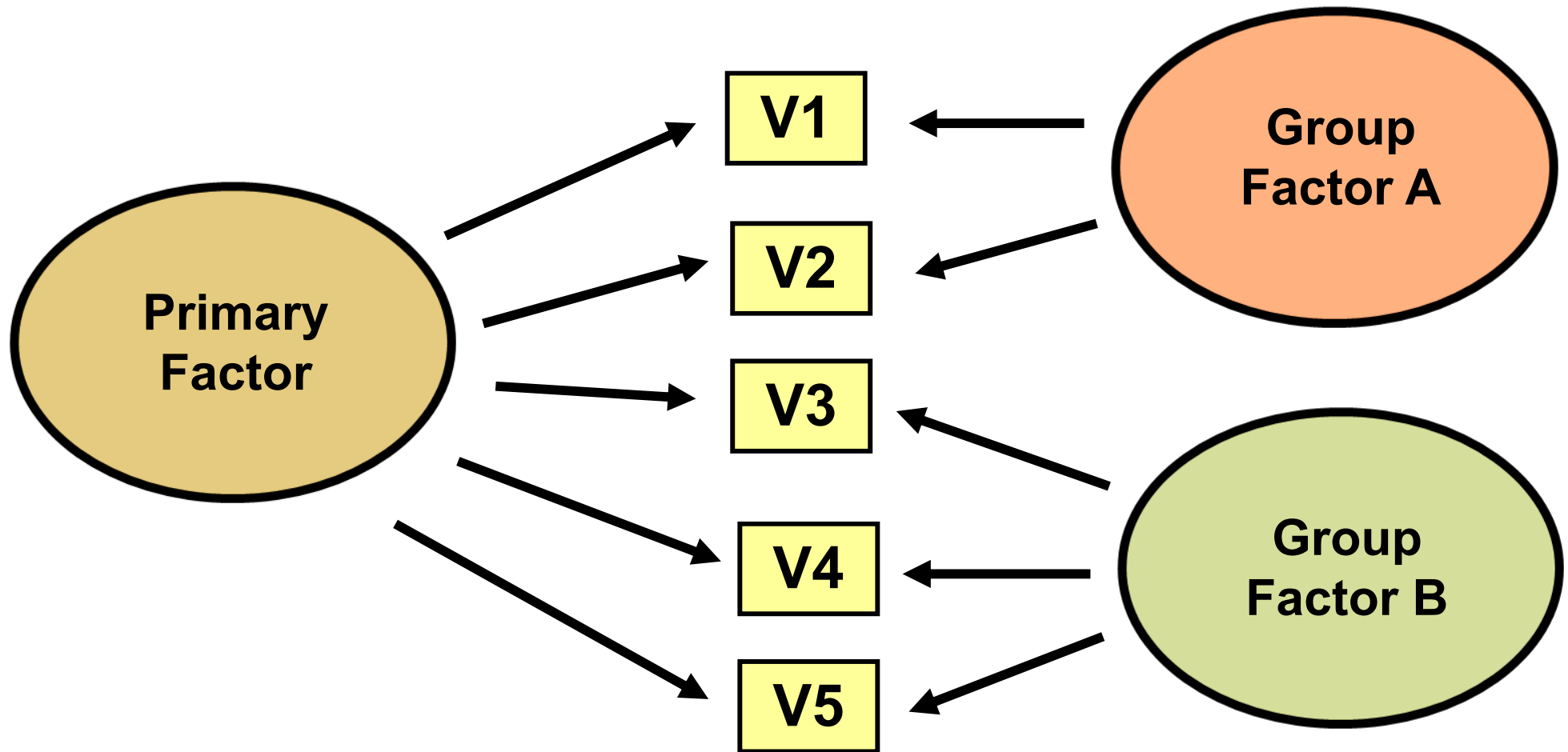


Simple structure

two latent variables and five manifest variables



The bi-factor model with five manifest variables and two group factors



Raw data matrices

Item data

(metrics may or may not differ)

	q1	q2	q3	q4	q5	q6	q7	...	qp
1	0	1	3	0	2	3	1	3	0
2
3
4
5
6
7
...
N									

Entries are 'scores' on items

Other multivariate data

(metrics will likely differ)

	v1	v2	v3	v4	v5	v6	v7	...	vp
1	16	126	0	1.5	0	-1	36	1	10
2
3
4
5
6
7
..
N									

Entries are values on variables

Association matrices

Variance-Covariance

	1	2	3	4	5	6	7	...	p
1	S^2								
2		S^2							
3			S^2						
4				S^2					
5					S^2				
6						S^2			
7							S^2		
...								S^2	
p									S^2

 = covariance

Correlation

	1	2	3	4	5	6	7	...	p
1	1								
2		1							
3			1						
4				1					
5					1				
6						1			
7							1		
...								1	
p									1

 = correlation coefficient

Specific matrices

$$\begin{bmatrix} 2 & 4 & 6 \\ 1 & 5 & 7 \\ 4 & 8 & 9 \end{bmatrix}$$

A

symmetric

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

B

diagonal

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

C

identity

Principal component model

Linear equation (in standard form)

$$z_j = a_j z_a + b_j z_b + u_j z_{uj}$$

$$z_k = a_k z_a + b_k z_b + u_k z_{uk}$$

z_j - standard score on item/variable j

z_k - standard score on item/variable k

z_a - standard score on component A

z_b - standard score on component B

a, b, u - loadings (weights) on the common and unique components

common and unique components are assumed to be mutually uncorrelated

Partitioning the total variance (of an item or variable) common factor and principal component models

$$1.00 = a^2 + b^2 + s^2 + e^2$$

Common factor
model

Communality (h^2)

specificity

error

Principal
component
model

Communality (h^2)

Uniqueness (U^2)

Reliability

Factor pattern matrix (F) and uniqueness vector (U²)

“Test” / Variable (p)		F1	F2	F3	h ²	U ²
1. Addition		.15	.01	.88	.79	.21
2. Cubes (3-D reasoning)		.89	.06	.07	.80	.20
3. Spelling		.03	.88	.08	.78	.22
4. Multiplication		.17	.18	.87	.82	.18
5. Flags (2-D reasoning)		.83	.01	.12	.70	.30
6. Vocabulary		.03	.85	.02	.72	.28
7. Block Counting		.77	.06	.17	.63	.37
8. Arithmetic reasoning		.46	.55	.36	.64	.36
9. Syllogisms		.59	.42	.24	.78	.22
Eigenvalue	$\Sigma(\text{load}^2)$	2.69	2.02	1.77		
Proportion covariance	$\frac{\Sigma(\text{load}^2)}{p}$.30	.22	.20		

From Thurstone LL (1938), *Primary Mental Abilities*.

Phi coefficients

- **Pearson product-moment correlations between dichotomous variables**
- **Influenced by two considerations**
 - **strength of relationship**
 - **differences in variable ‘means’ (skewnesses or “marginal splits”)**
- **May lead to spurious or “difficulty” factors, which are artifacts resulting from nonlinear regression of items on factors**

**Maximum Pearson product-moment correlation
(r , phi coefficient) as a function of the “marginal splits”
of two dichotomous variables**

↓ V1 V2 →	.10/.90	.20/.80	.50/.50	.70/.30	.90/.10
.10/.90	1.00				
.20/.80	.66	1.00			
.50/.50	.33	.50	1.00		
.70/.30	.22	.33	.66	1.00	
.90/.10	.11	.17	.33	.51	1.00

Approaches to the problem of “attenuation of max phi”

- **Phi/phi_{max} adjusted on basis of max r**
 - Not a true correlation coefficient, association matrix problematic
- **Use tetrachoric correlation coefficients in the matrix**
 - Estimates with assumption that variables “normally distributed in underlying form”; often inappropriate; can inflate coefficients
- **Miniscales, radial parcels; averaging procedures**
 - Can yield spurious factors
- **G-index of agreement**
 - = split via reproduced scores for each item, extended data matrix
- **Estimate r among latent mvnl vars via dichotomies, then determine factor structure (e.g., LISCOMP)**

Empirical comparisons show none to be superior.

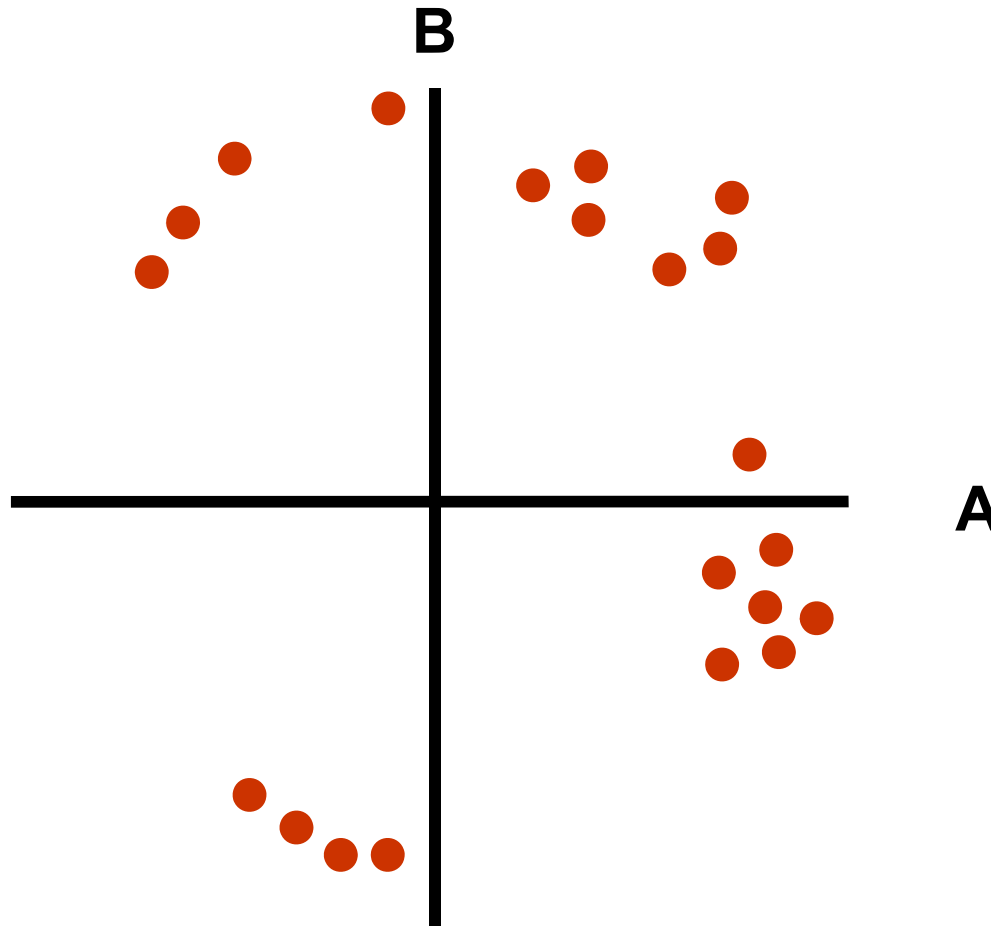
Elements contributing to a stable solution

- Subject sampling (n)
- Variable sampling (p)
- Subject-to-variable ratio (n:p)
- High average magnitude of loadings (saturation)
- Saturation / ($n^{1/2}$)
- Well-identified components (several salient variables)
- Striking a balance between “over-” and “under-extraction”
- Exploratory and confirmatory methods

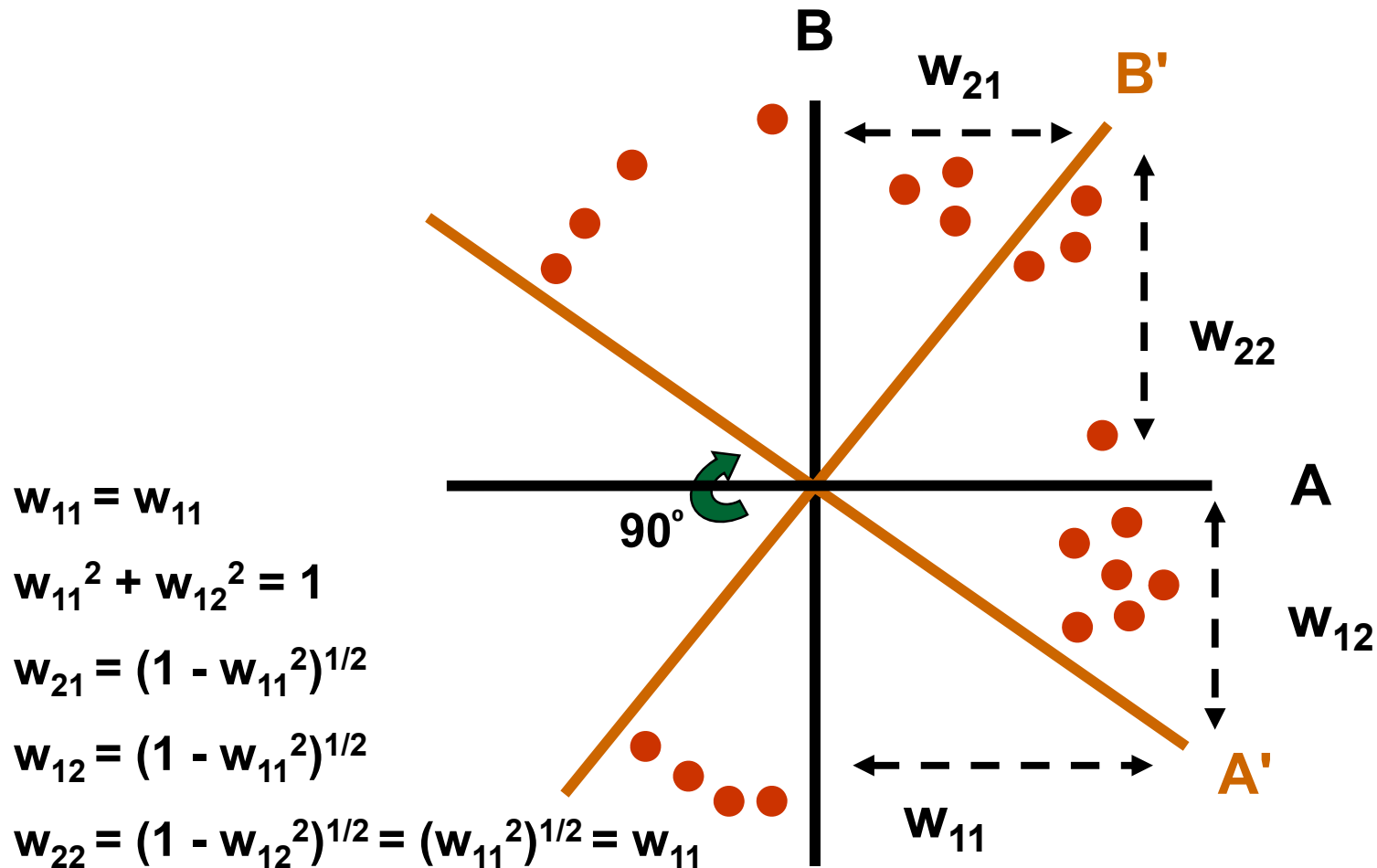
e.g., Velicer & Fava. *Psychological Methods*. 1998;3:231-251.

Wood, Tataryn & Gorsuch. *Psychological Methods*. 1996;1:354-365.

Two unrotated components (21 items)



Orthogonal rotation for simple structure



Adapted from RL Gorsuch, *Factor Analysis*, Erlbaum, 1983. pp 219-220.

Vanderbilt Transplant Center HRQOL survey battery and assessment schedule (effective January, 2002)

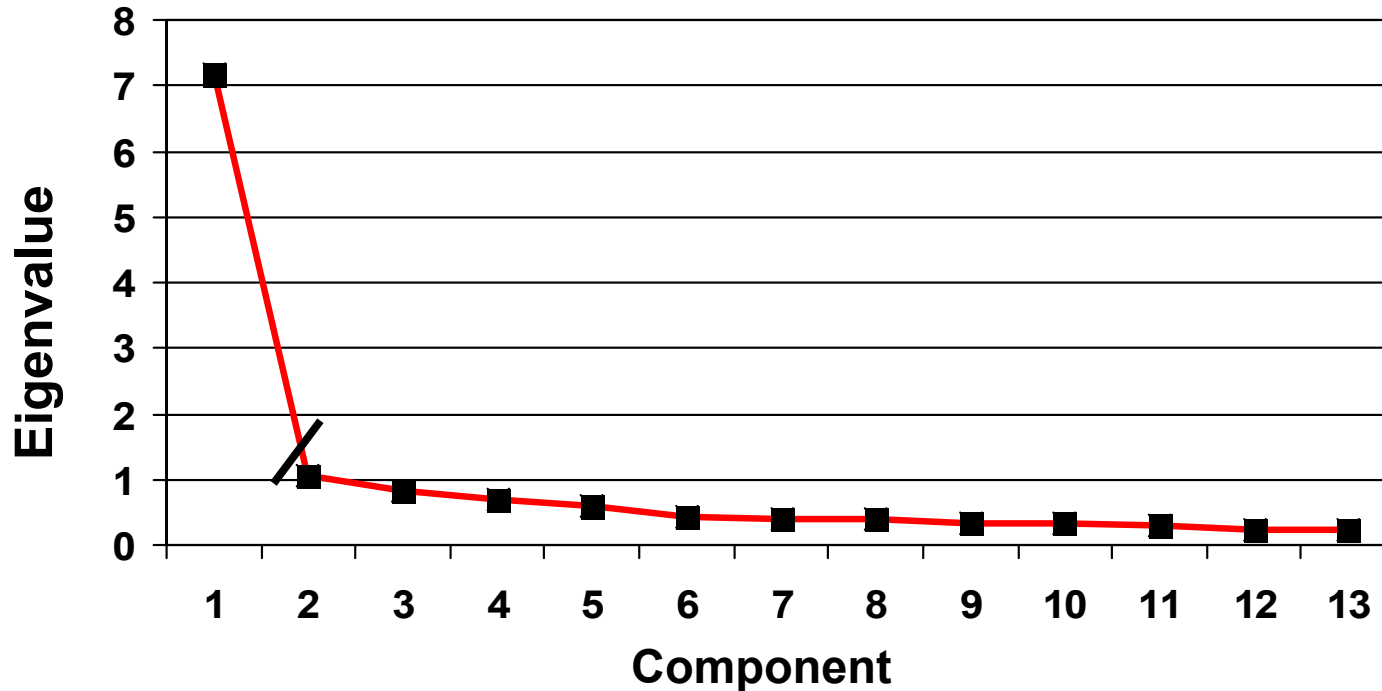
Survey Instrument	(# items)	Eval	Listed q 6 M	6 Hr Pre	1 M Post	3 M Post	6 M Post	Ann Post
SF-36®	(36)	X	X		X	X	X	X
Beck Anxiety Inventory	(21)	X	X		X	X	X	X
CES-D Depression Scale	(20)	X	X		X	X	X	X
Satisfaction Inventory	(16)	X	X		X	X	X	X
Overall Health (VAS)	(1)	X	X		X	X	X	X
Employment	(20 pre, 17 post)	X					X	X
EQ-5D	(5)	X	X		X	X	X	X
Symptom survey	(pending)	X	X		X	X	X	X
Functional Performance	(1)	X	X	X	X	X	X	X

PAIS discontinued January, 2002
 EQ-5D added July, 2006
 Symptom survey being developed


Listing **Transplant**

Scree plot of eigenvalues

13 core Vanderbilt Transplant Center Patient Satisfaction Inventory (VTCPSI) items

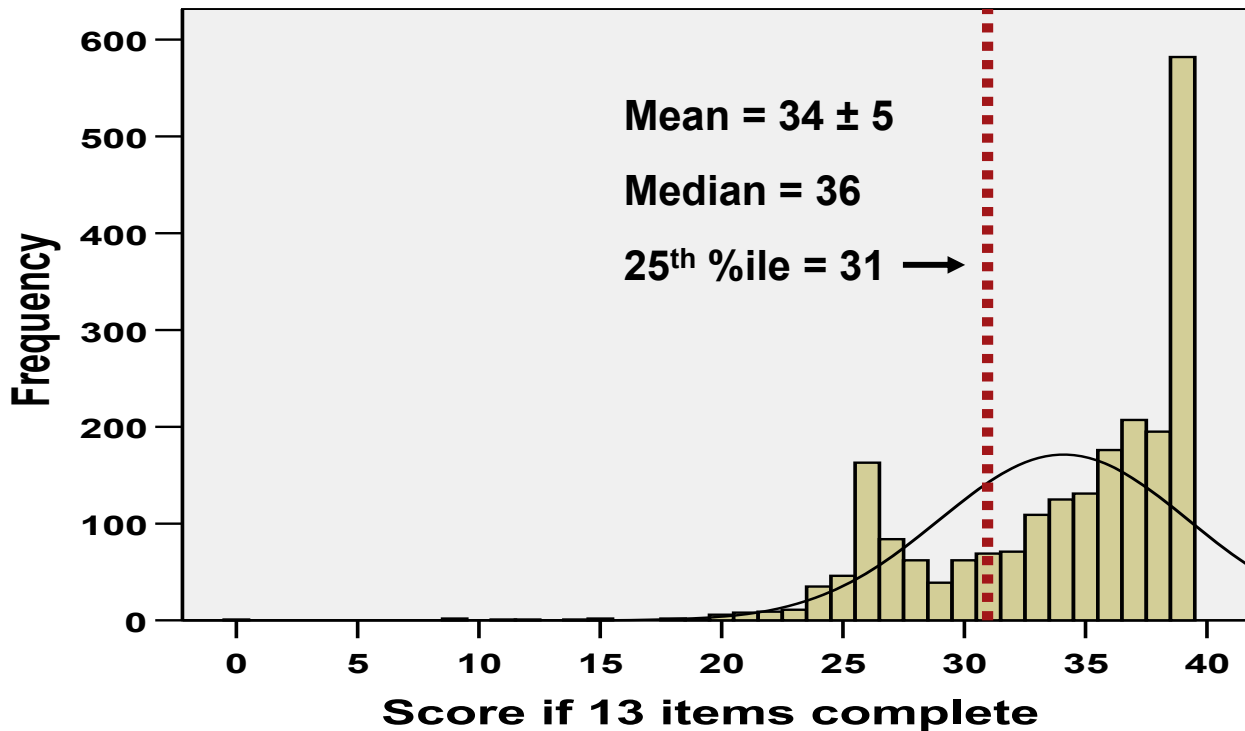


The first component accounts for 56% of the covariance.
The second component accounts for only an additional 9%.

Summary parameters - VTCPSI

Time point	N	13 items complete	n:p	KMO	Eigen-val 1, 2	Comp 1 (% var)	Min load	Sat-uration	Coeff α
Eval	546	66%	28	.91	7.4 1.2	57%	.53	.75	.92
Wait	514	82%	32	.92	7.4 1.3	57%	.61	.75	.93
1 M	229	81%	14	.90	7.3 1.0	56%	.56	.74	.92
3 M	177	89%	12	.91	7.0 1.1	54%	.54	.73	.92
6 M	181	89%	12	.91	6.9 1.3	53%	.54	.72	.91
Ann	1061	86%	70	.95	7.2 1.2	56%	.51	.74	.92
Total	2708	81%	169	.94	7.2 1.2	56%	.55	.74	.92

Distribution of satisfaction inventory scores (all respondents and monitoring points)



Unpublished data: Quantitative Dizziness Questionnaire headache scale

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.618	.615	6

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
G3	8.17	19.747	.124	.099	.649
G5	9.11	17.361	.292	.088	.597
H1	9.21	15.460	.449	.405	.532
H2	9.60	16.221	.454	.439	.535
H6	9.71	15.918	.328	.151	.587
H7	9.47	15.798	.469	.240	.527

Measuring self-report symptoms of depression in organ transplant patients: validation of a multidimensional model of CES-D item data

- Organ transplant candidates and recipients may present with signs and symptoms of physical and/or psychogenic origin that could be interpreted as depression.
- Sometimes difficult to differentiate depression from fatigue
- Corticosteroid-based immunosuppression is associated with symptoms of depression and anxiety.
(Yehuda et al. *Biological Psychiatry* 1993 v34, MacNaughton et al. *Clinical Transplantation* 1998 v12)
- Has direct bearing on the interpretation and scoring of items (and scales) that focus on somatic symptoms in the transplant setting

Random sample by symptom severity group (respondents with complete CES-D item data)

			Random Sample		Total
			training	test	
CES-D Symptoms (2 levels)	none/mild	Count	454	423	877
		% within CES-D Symptoms (2 levels)	51.8%	48.2%	100.0%
		% within Random Sample	69.7%	69.0%	69.4%
	mod/severe	Count	197	190	387
		% within CES-D Symptoms (2 levels)	50.9%	49.1%	100.0%
		% within Random Sample	30.3%	31.0%	30.6%
Total	Count	651	613	1264	
	% within CES-D Symptoms (2 levels)	51.5%	48.5%	100.0%	
	% within Random Sample	100.0%	100.0%	100.0%	

P = 0.824

PCA with orthogonal (varimax) rotation

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.548	37.740	37.740	7.548	37.740	37.740	4.644	23.222	23.222
2	1.915	9.576	47.317	1.915	9.576	47.317	3.508	17.540	40.762
3	1.332	6.661	53.978	1.332	6.661	53.978	2.643	13.217	53.978
4	.996	4.982	58.960						
5	.806	4.031	62.992						
6	.771	3.854	66.846						
7	.738	3.688	70.533						
8	.673	3.365	73.898						
9	.620	3.101	76.999						
10	.590	2.952	79.951						
11	.559	2.796	82.747						
12	.523	2.616	85.363						
13	.488	2.439	87.801						
14	.461	2.306	90.107						
15	.442	2.208	92.315						
16	.369	1.845	94.160						
17	.333	1.665	95.825						
18	.314	1.571	97.395						
19	.295	1.473	98.869						
20	.226	1.131	100.000						

Component matrices (training sample)

Unrotated (54%)	Component		
	1	2	3
Bothered by things	.700	.129	-.050
Poor appetite	.523	.134	-.340
Can't shake blues	.798	.080	.003
Just as good as others	-.327	.628	-.067
Trouble concentrating	.646	.199	-.175
Felt depressed	.812	.088	-.008
Everything an effort	.614	.277	-.320
Hopeful about future	-.399	.664	.101
Life a failure	.585	-.041	.344
Fearful	.680	.093	.224
Restless sleep	.570	.182	-.341
Happy	-.554	.598	.169
Talked less	.564	.057	-.007
Lonely	.762	.022	.175
People unfriendly	.372	.069	.486
Enjoyed life	-.480	.653	.161
Crying spells	.557	.145	.238
Felt sad	.813	.091	.108
People disliked me	.520	-.008	.518
Could not get going	.693	.222	-.351

Varimax (54%)	Component		
	1	2	3
Bothered by things	.582	.387	-.142
Poor appetite	.625	.051	-.115
Can't shake blues	.604	.483	-.215
Just as good as others	.039	-.179	.687
Trouble concentrating	.642	.263	-.076
Felt depressed	.623	.483	-.214
Everything an effort	.733	.137	-.014
Hopeful about future	-.097	-.084	.771
Life a failure	.206	.617	-.200
Fearful	.395	.590	-.128
Restless sleep	.678	.083	-.088
Happy	-.274	-.128	.776
Talked less	.432	.334	-.153
Lonely	.456	.592	-.231
People unfriendly	.011	.615	-.001
Enjoyed life	-.195	-.085	.799
Crying spells	.317	.535	-.035
Felt sad	.557	.576	-.195
People disliked me	.070	.721	-.121
Could not get going	.787	.152	-.098

Orthogonal rotation (varimax)

Varimax (54%)	Component		
	1	2	3
Bothered by things	.582	.387	
Poor appetite	.625		
Can't shake blues	.604	.483	
Just as good as others			.687
Trouble concentrating	.642		
Felt depressed	.623	.483	
Everything an effort	.733		
Hopeful about future			.771
Life a failure		.617	
Fearful	.395	.590	
Restless sleep	.678		
Happy			.776
Talked less	.432	.334	
Lonely	.456	.592	
People unfriendly		.615	
Enjoyed life			.799
Crying spells	.317	.535	
Felt sad	.557	.576	
People disliked me		.721	
Could not get going	.787		

Varimax (54%)	Component		
	1	2	3
Bothered by things	.582	.387	
Poor appetite	.625		
Can't shake blues	.604	.483	
Just as good as others			.687
Trouble concentrating	.642		
Felt depressed	.623	.483	
Everything an effort	.733		
Hopeful about future			.771
Life a failure		.617	
Fearful	.395	.590	
Restless sleep	.678		
Happy			.776
Talked less	.432	.334	
Lonely	.456	.592	
People unfriendly		.615	
Enjoyed life			.799
Crying spells	.317	.535	
Felt sad	.557	.576	
People disliked me		.721	
Could not get going	.787		

Two- and one-component solutions

Varimax (47%)	Component	
	1	2
Bothered by things	.695	
Poor appetite	.534	
Can't shake blues	.766	
Just as good as others		.705
Trouble concentrating	.673	
Felt depressed	.781	
Everything an effort	.674	
Hopeful about future		.768
Life a failure	.523	
Fearful	.662	
Restless sleep	.596	
Happy		.768
Talked less	.541	
Lonely	.710	
People unfriendly	.369	
Enjoyed life		.789
Crying spells	.569	
Felt sad	.783	
People disliked me	.476	
Could not get going	.724	

No rotation (38%)	Component
	1
Bothered by things	.700
Poor appetite	.523
Can't shake blues	.798
Just as good as others	-.327
Trouble concentrating	.646
Felt depressed	.812
Everything an effort	.614
Hopeful about future	-.399
Life a failure	.585
Fearful	.680
Restless sleep	.570
Happy	-.554
Talked less	.564
Lonely	.762
People unfriendly	.372
Enjoyed life	-.480
Crying spells	.557
Felt sad	.813
People disliked me	.520
Could not get going	.693

Cumulative proportion of variance by sample

Component(s) ↓ Sample →	Training	Test	Optimism	Total
1	0.37740	0.35285	0.02455	0.36523
1 & 2	0.47317	0.44275	0.03042	0.45785
1 - 3	0.53978	0.51822	0.02156	0.52824

Cumulative proportion of variance by model (in the test sample)

Latent Dimension	Model A (PCA)	Model B (Bi-factor)	Model C
General		0.41455	0.41777
Somatic symptoms	0.35285	0.06033	0.05883
Positive affect	0.08990	0.09153	0.09156
Isolation	0.07547	0.03645	
Total	0.51882	0.60286	0.56816

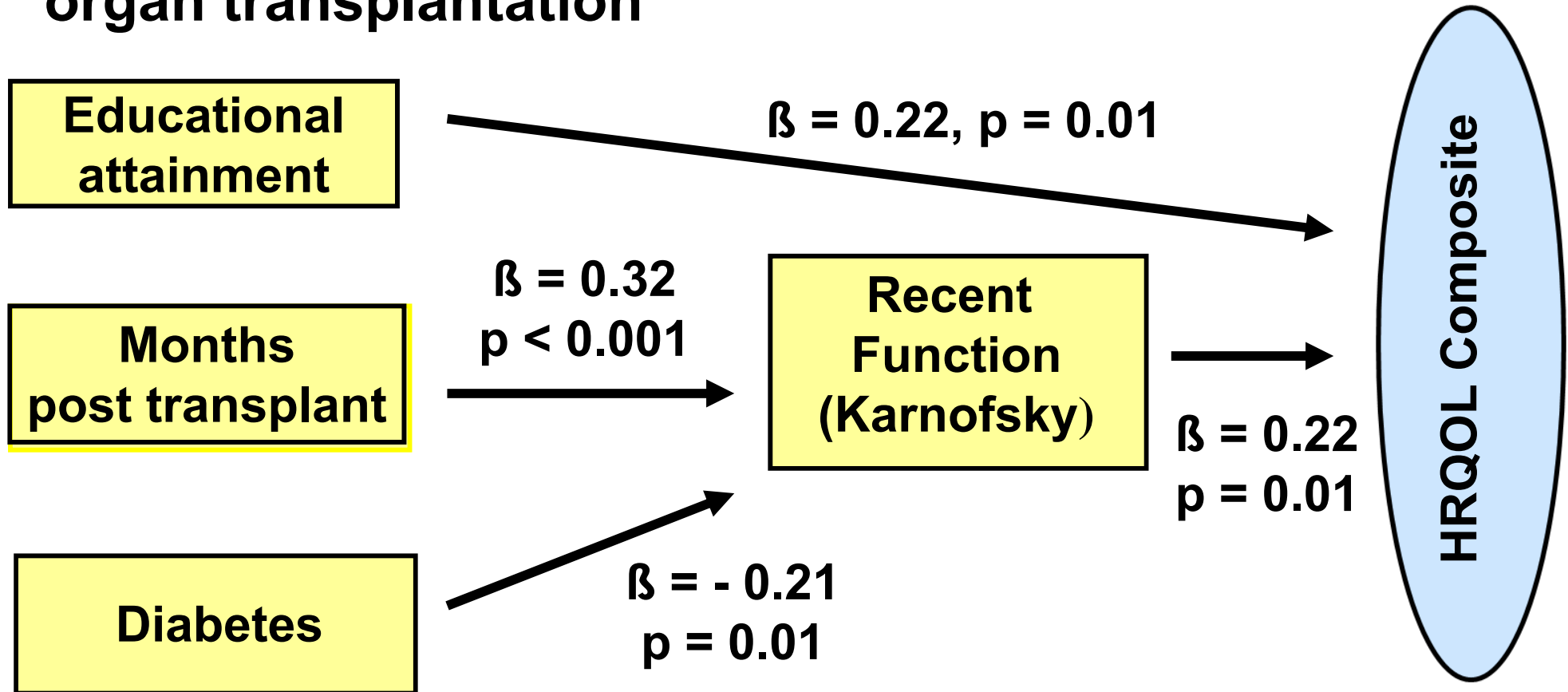
Diff $X^2_6 = 28.97$ ($p < 0.001$)

Health-related quality of life

2nd order principal component analysis

SF-36 Scale	Loading	PAIS Scale	Loading
Physical Function	.76	Health Care Orientation	.42
Physical Role	.74	Vocational	.65
Bodily Pain	.70	Domestic	.79
General Health	.75	Sexual Relations	.51
Vitality	.80	Family	.58
Social Function	.81	Social	.77
Emotional Function	.67	Psychological Distress	.76
Mental Health	.78	Cum prop variance (c1)	50%

PCA-derived composite outcome measure in a multivariate model of variables affecting health-related quality of life (HRQOL) after organ transplantation



PCA applications in microarray data analysis

two examples

Wang A and Gehan EA.

Gene selection for microarray data analysis using principal component analysis. *Statist Med.* 2005; 24:2069-2087.

Chen X, Wang L, Smith JD, Zhang B.

Supervised principal component analysis for gene set enrichment of microarray data analysis with continuous or survival outcomes. *Bioinformatics.* Advanced access Aug 27, 2008.