LONGITUDINAL MEASUREMENT INVARIANCE FOR A MARITAL SATISFACTION INSTRUMENT: WHICH IS UNSTABLE: THE CONSTRUCT OR THE INSTRUMENT?

Antonio Olmos*, Susan Hutchinson#, Scott Stanley*, Galena Rhoades*

American Evaluation Association Conference
Minneapolis, MN. October 2012

* University of Denver
# University of Northern Colorado

Part of this research was supported by National Institute of Mental Health (NIMH), Division of Services and Intervention Research, Adult and Geriatric Treatment and Prevention Branch Grant 5-RO1-MH35525-12, “The Long-Term Effects of Premarital Intervention.”

It started with some strange reliabilities…

- Reliabilities in the early waves were abysmal (0.4)
- But as we collected more data over time, a funny thing happened:
  - Reliabilities started to improve
  - From follow-up wave 3 and on, reliabilities were 0.7

This pattern of increasing reliability over time raised the question:

Was the Locke-Wallace Relationship Adjustment Test measuring the same marital satisfaction trait across time?

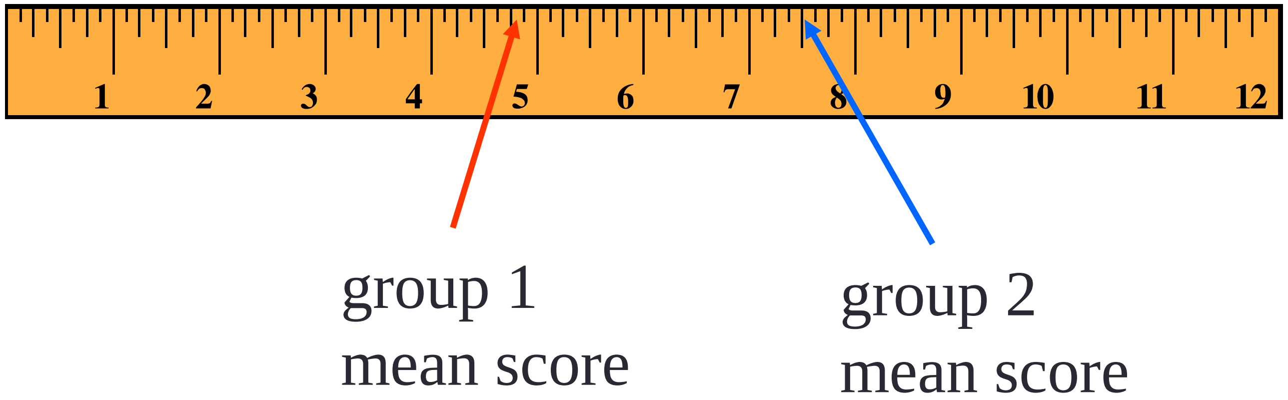
And if not …. what were the implications for examining change in marital satisfaction levels?

Further, do we really need to concerned given that the Locke-Wallace has been “well-established”?

Concept of Measurement Invariance

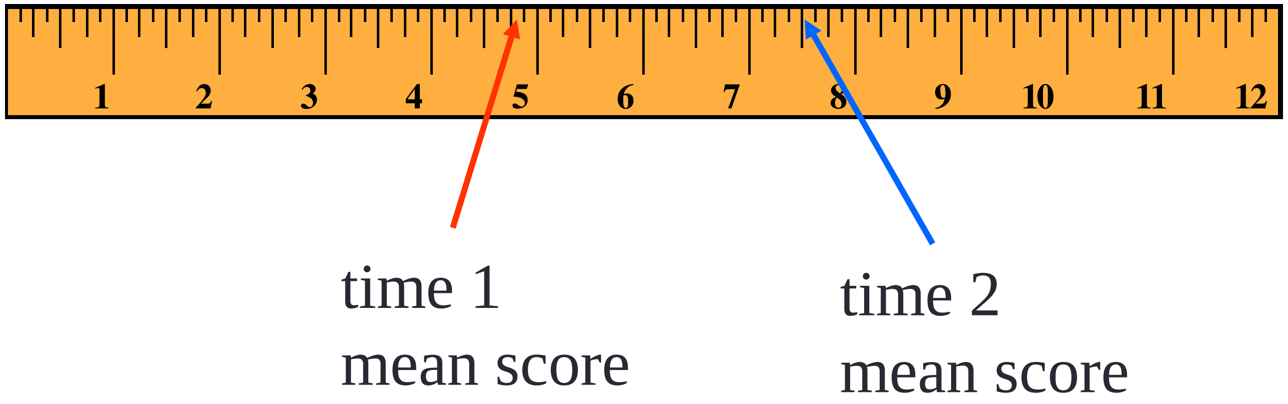
- The same measurement model (e.g., factor model, IRT model, etc.) holds across different populations or time periods
- Items mean the same thing to all respondents
- Participants understand and use measurement scales or response options in the same way, e.g.,
  - An option of “rarely” represents the same quantity for all respondents

Measurement invariance – cont'd



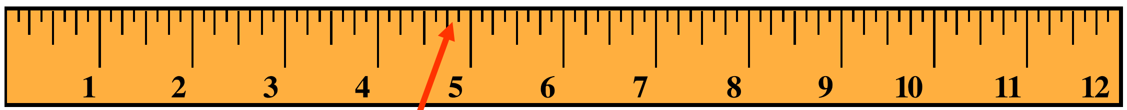
When a measure produces equivalent scores it is analogous to placing scores along the same linear continuum, allowing meaningful comparisons between groups or …

Measurement invariance – cont'd



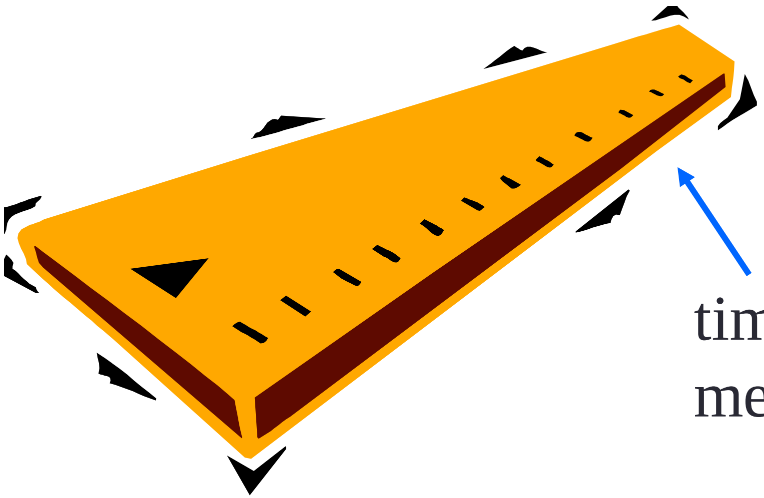
When a measure produces equivalent scores it is analogous to placing scores along the same linear continuum, allowing meaningful comparisons between groups or **across time**

Measurement non-invariance illustrated

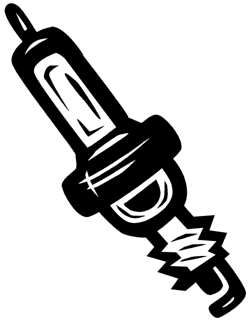


Trait "A"

time 1
mean score



Trait "B"



time 2
mean score

Why is longitudinal measurement invariance important?

- Measurement invariance is a validity issue.
- Without evidence of equivalent measurement, scores across time cannot be considered equally valid, i.e.,
  - The absence of equivalence compromises score interpretation for at least some participants across the waves of data collection
- Without evidence of equivalence, tests of mean differences cannot be unambiguously interpreted
  - Cannot know if apparent mean differences reflect change in level of the trait, change in nature of the trait, or merely a measurement artifact

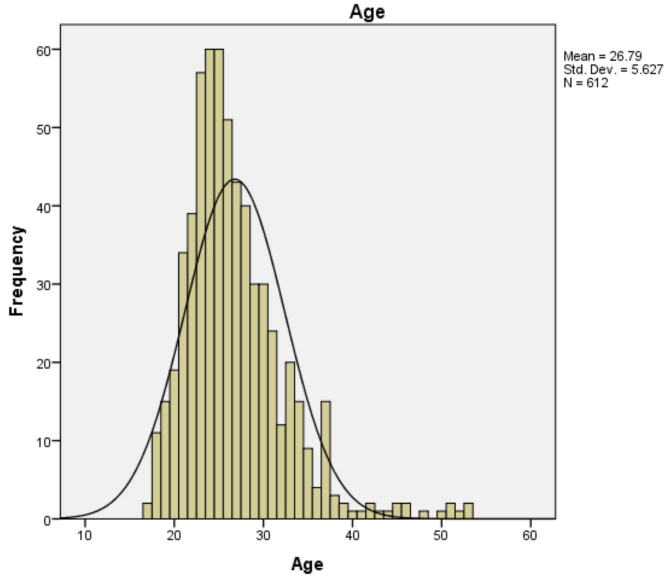
Longitudinal invariance illustrated using the Locke-Wallace

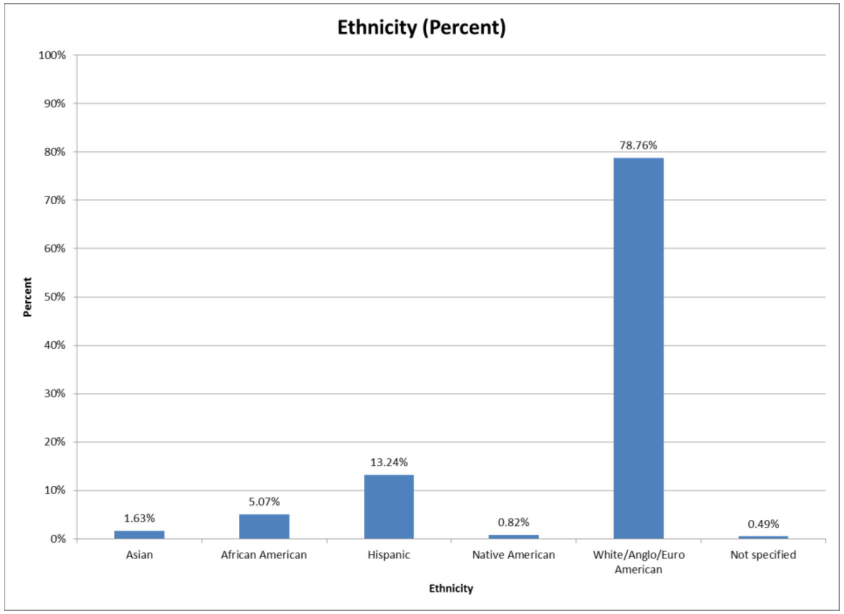
- Study designed to identify a non-convenience community sample representative of the couples marrying in religious organizations (ROs) in Denver.
- Sample of 105 large ROs,
  - Invited couples seeking marriage at their organization to participate in the study (for details, see Stanley et al., 2001).
  - 306 couples from recruited ROs participated in 3 conditions

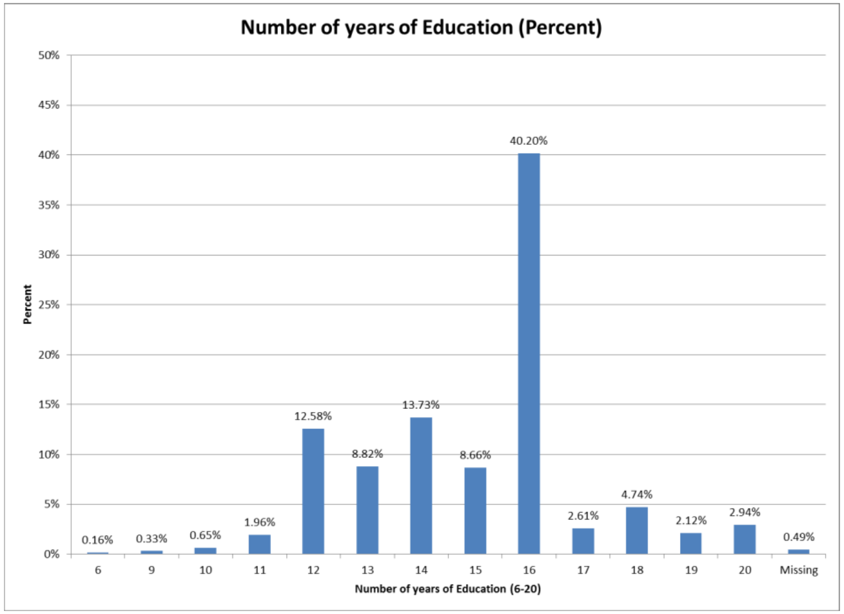
Locke Wallace

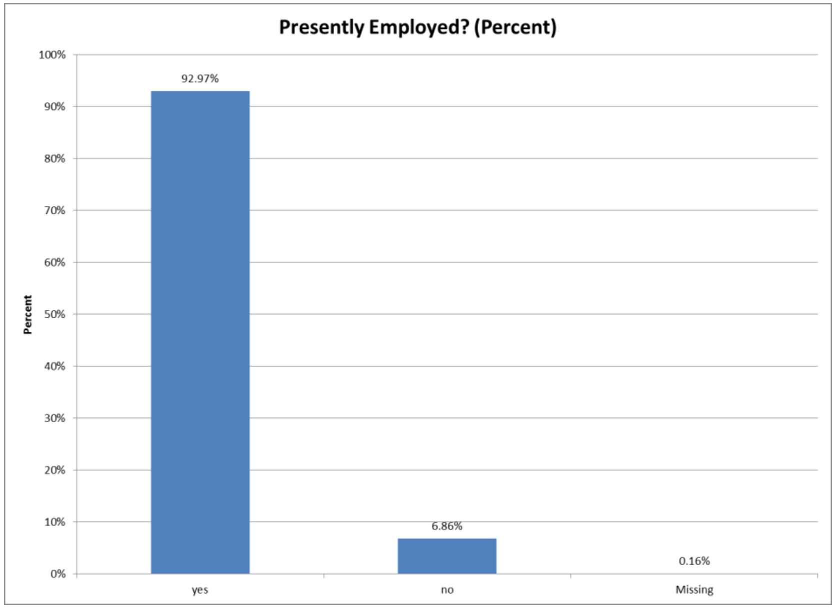
- Developed in 1959
- 16 items purported to be unidimensional
- “Strange” item weighting to maximize discriminative power
- Sample items
  - Handling family finances
  - Matters of recreation
  - Affection
  - Do you ever wish you had not married
  - In leisure time do you prefer to….
  - Global happiness item

Demographics

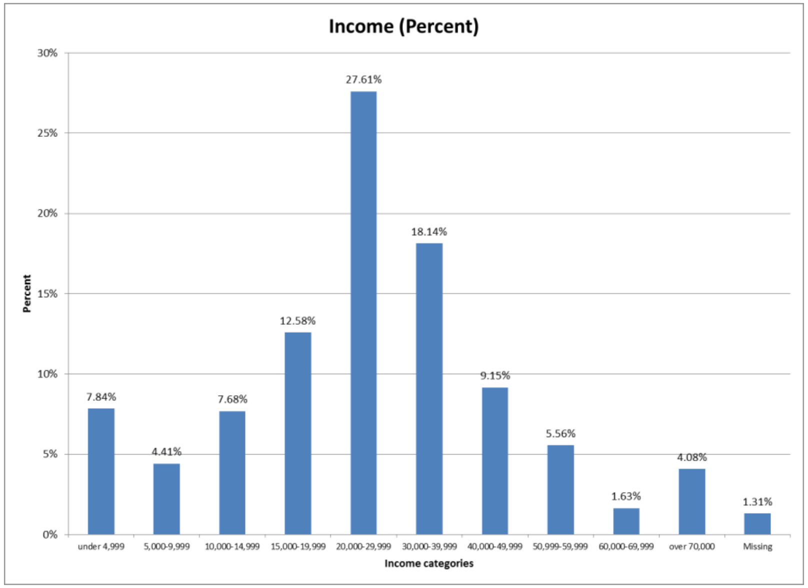


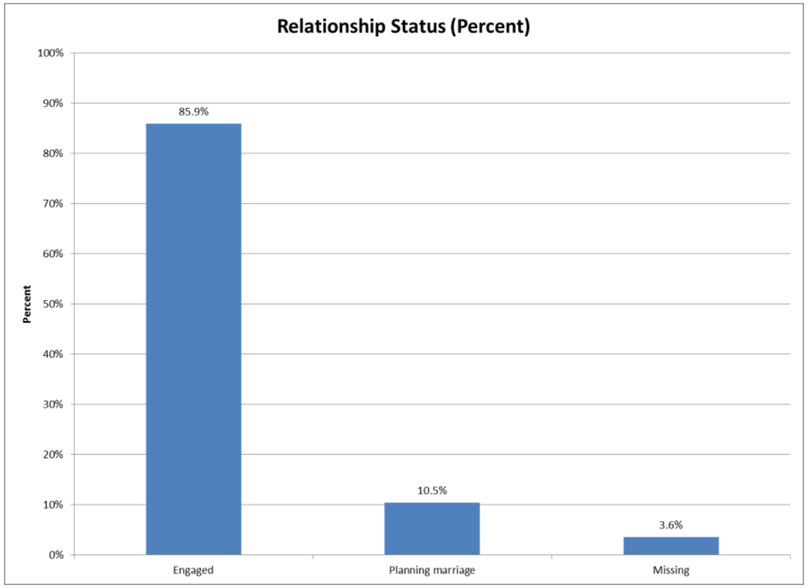


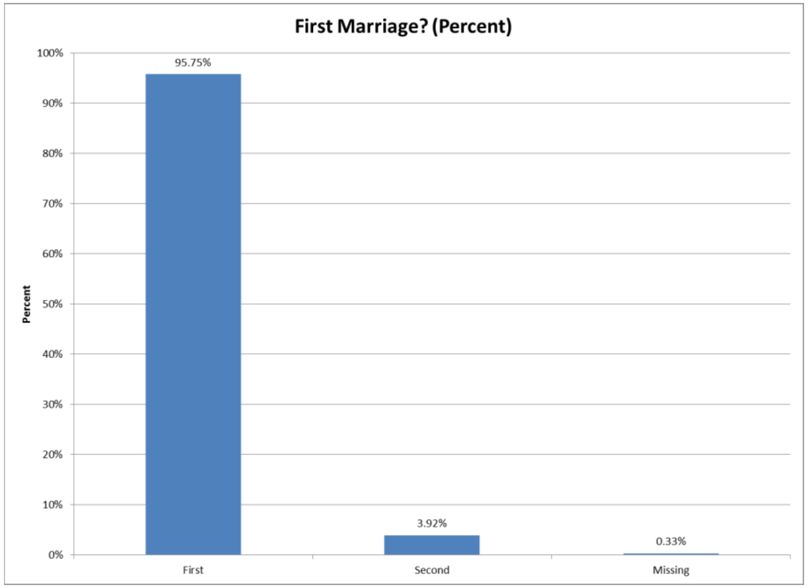


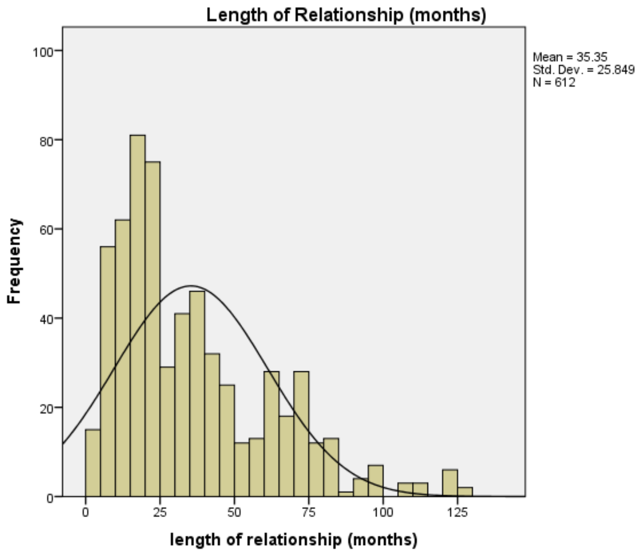


Demographics (cont’d)

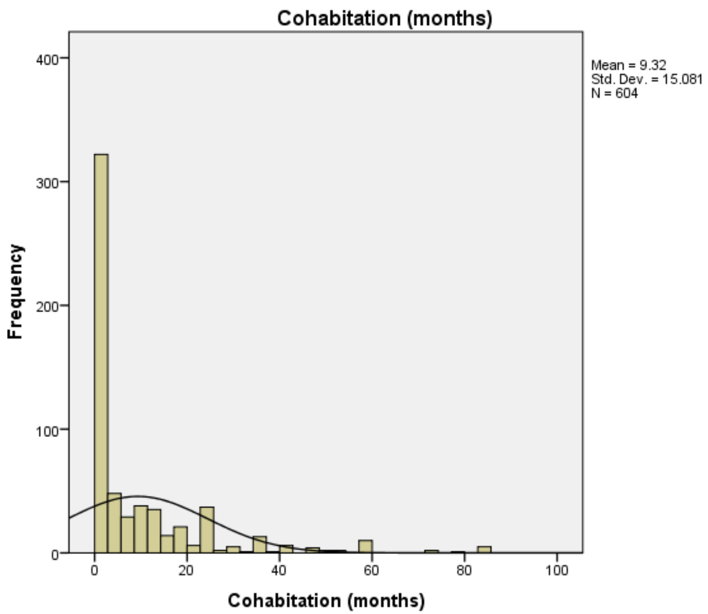


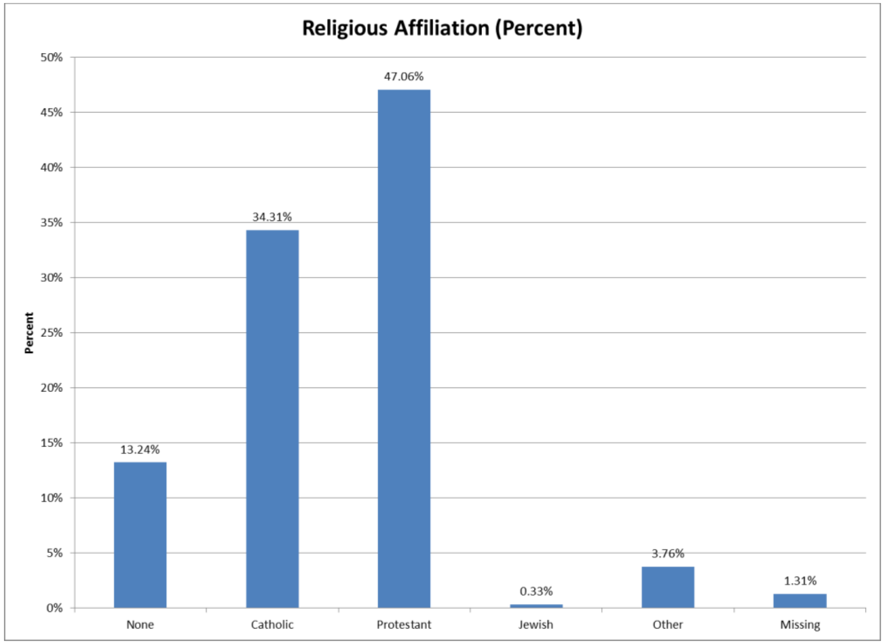




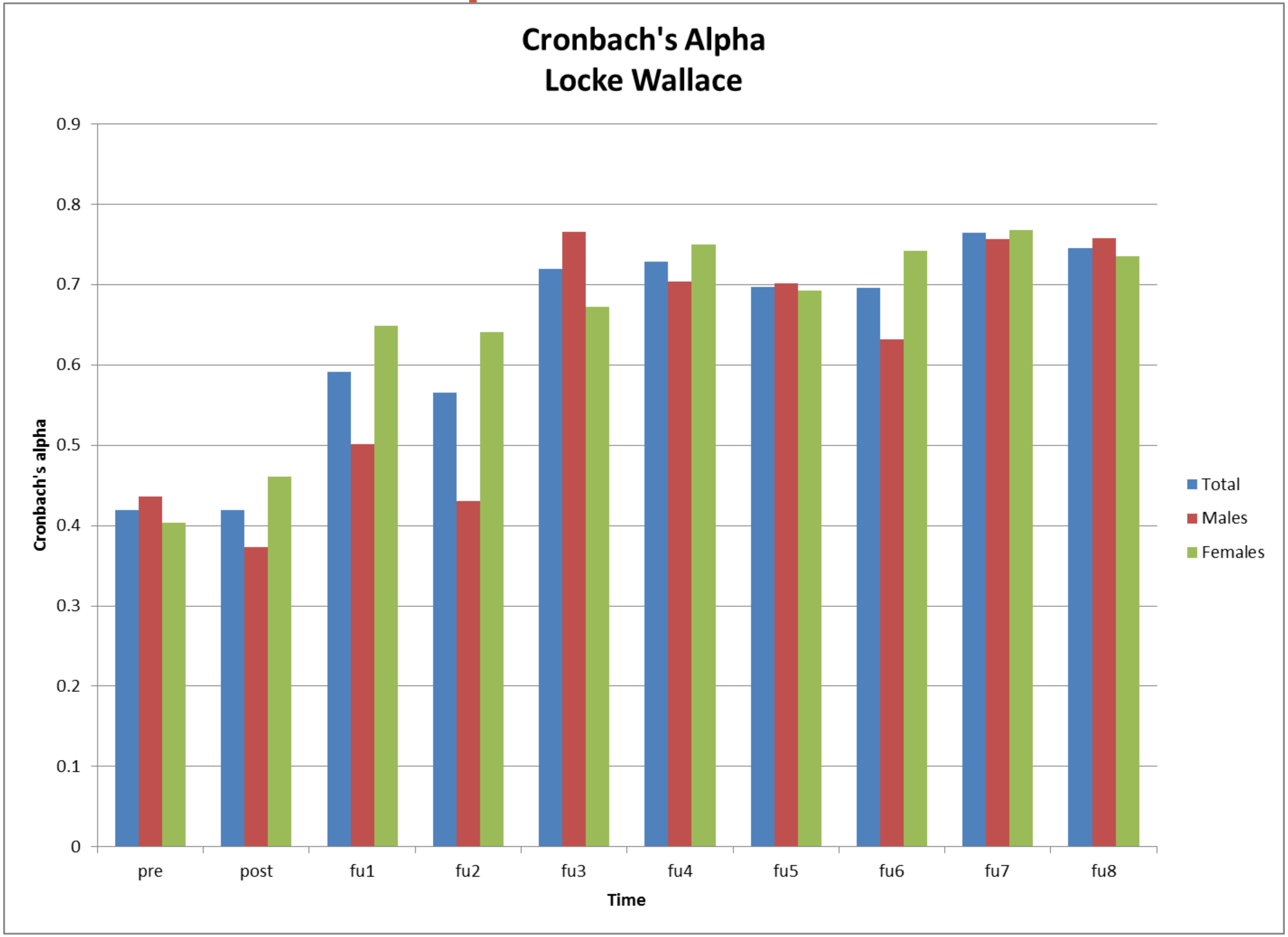


Demographics (cont'd)





Cronbach's alpha over time



Rank-Order Correlations (total score)

Males and Females

a. Listwise N = 44

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	.480									
FU1	.477	.621								
FU2	.061	.145	.393							
FU3	.108	.324	.502	.509						
FU4	.242	.333	.406	.471	.794					
FU5	.265	.298	.515	.488	.710	.741				
FU6	.325	.316	.412	.301	.571	.693	.596			
FU7	.290	.247	.403	.422	.595	.667	.634	.514		
FU8	.384	.404	.477	.289	.449	.562	.627	.652	.694	

Pairwise

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	0.524									
FU1	0.473	0.525								
FU2	0.297	0.367	0.477							
FU3	0.201	0.414	0.433	0.551						
FU4	0.268	0.375	0.489	0.547	0.636					
FU5	0.292	0.395	0.511	0.523	0.589	0.716				
FU6	0.35	0.363	0.454	0.484	0.495	0.705	0.641			
FU7	0.392	0.402	0.457	0.452	0.47	0.666	0.692	0.628		
FU8	0.308	0.329	0.431	0.406	0.401	0.619	0.646	0.61	0.671	

Rank-order correlations (total score) by Gender

Males

	a. gender = male			b. Listwise N = 25						
	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	.456									
FU1	.300	.477								
FU2	-.172	-.033	.366							
FU3	-.069	.076	.426	.483						
FU4	.181	.207	.444	.494	.810					
FU5	.166	.230	.476	.392	.755	.810				
FU6	.272	.162	.378	.163	.436	.606	.542			
FU7	.282	.207	.586	.266	.612	.664	.562	.385		
FU8	.377	.455	.636	.138	.444	.604	.563	.662	.666	

pairwise

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	0.555									
FU1	0.368	0.483								
FU2	.128	0.318	0.438							
FU3	.113	0.424	0.459	0.647						
FU4	0.25	0.299	0.562	0.608	0.721					
FU5	.183	0.371	0.568	0.469	0.609	0.739				
FU6	0.305	0.352	0.471	0.468	0.487	0.721	0.646			
FU7	0.348	0.408	0.532	0.395	0.541	0.692	0.682	0.596		
FU8	0.363	0.33	0.542	0.435	0.562	0.632	0.627	0.637	0.692	

Females

	a. gender = female			b. Listwise N = 19						
	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	.443									
FU1	.737	.761								
FU2	.374	.338	.376							
FU3	.269	.644	.539	.511						
FU4	.275	.437	.234	.364	.690					
FU5	.458	.296	.443	.611	.527	.516				
FU6	.281	.530	.408	.451	.706	.764	.633			
FU7	.186	.163	.062	.653	.473	.639	.692	.602		
FU8	.304	.174	.155	.524	.419	.422	.738	.574	.751	

pairwise

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE										
POST	0.49									
FU1	0.575	0.559								
FU2	0.459	0.417	0.527							
FU3	0.29	0.412	0.38	0.456						
FU4	0.286	0.459	0.402	0.485	0.528					
FU5	0.405	0.43	0.457	0.583	0.557	0.699				
FU6	0.395	0.377	0.454	0.525	0.513	0.686	0.634			
FU7	0.425	0.421	0.397	0.489	0.386	0.635	0.683	0.64		
FU8	0.279	0.356	0.326	0.359	0.216	0.605	0.662	0.583	0.647	

DIF Fit Summary

Males

PERSON CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	ITEM Number	Name
10	20.5474	9	.0148	.3081	-1.9088	1	LW_1A
10	2.4038	9	.9833	.0382	-4.0632	2	LW_2A
10	10.1687	9	.3369	.1350	-2.9423	3	LW_3A
10	31.4540	9	.0002	.4706	-1.2567	4	LW_4A
10	5.4888	9	.7898	.1007	-3.2460	5	LW_5A
10	74.5702	9	.0000	1.1536	.4675	6	LW_6A
10	5.0737	9	.8278	.0741	-3.5339	7	LW_7A
10	19.7648	9	.0194	.2923	-1.9834	8	LW_8A
10	5.2081	9	.8158	.0702	-3.5810	9	LW_9A
10	5.2147	9	.8152	.0514	-3.8413	10	LW_10A
10	53.2598	9	.0000	.7517	-.4204	11	LW_11A
10	9.5801	9	.3855	.1222	-3.0490	12	LW_14A
10	2.3132	9	.9855	.0545	-3.7942	13	LW_15A
10	7.0296	9	.6340	.0936	-3.3176	14	LW_16A
10	36.7132	9	.0000	.5178	-1.0963	15	LWLEI

Females

PERSON CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	ITEM Number	Name
10	29.6951	9	.0005	.4500	-1.3300	1	LW_1A
10	4.8383	9	.8482	.0472	-3.9073	2	LW_2A
10	20.6505	9	.0143	.3515	-1.7155	3	LW_3A
10	45.6318	9	.0000	.6937	-.5731	4	LW_4A
10	10.2532	9	.3303	.1466	-2.8511	5	LW_5A
10	54.5388	9	.0000	.9261	-.0037	6	LW_6A
10	3.3088	9	.9508	.0345	-4.1343	7	LW_7A
10	7.3006	9	.6058	.0850	-3.4090	8	LW_8A
10	6.3276	9	.7067	.0754	-3.5177	9	LW_9A
10	10.7635	9	.2922	.1424	-2.8832	10	LW_10A
10	38.1664	9	.0000	.5420	-1.0180	11	LW_11A
10	5.5224	9	.7866	.0650	-3.6474	12	LW_14A
10	15.3456	9	.0818	.2120	-2.4123	13	LW_15A
10	7.6822	9	.5664	.1134	-3.1260	14	LW_16A
10	46.9066	9	.0000	.6683	-.6428	15	LWLEI

Items with Differential functioning across waves (pairwise comparisons)

FEMALES

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE				6	4,6	6,11	4,6,11,L	4,6	1,4,6	4,6
POST						11,L	11,L	L		
FU1										
FU2										
FU3										
FU4										
FU5	3									
FU6	3									
FU7	3									
FU8	3									

MALES

	PRE	POST	FU1	FU2	FU3	FU4	FU5	FU6	FU7	FU8
PRE					4,6,11	4,6,11,L	4,6,L	6,11	6,11	6,11
POST					6,11	6	6,11,L	6,11	6,11	6,11
FU1	8						L			
FU2	8									
FU3	8									
FU4										
FU5										
FU6										
FU7										
FU8										

Upper diagonal: $p < 0.0005$
Lower diagonal: 0.5 Logits of difference

Multiple Groups Confirmatory Factor Analysis (CFA)

- Conducted pairwise, between each adjacent wave of data, e.g., pre vs post, post vs follow-up 1, etc.
- Series of tests to assess stability of factor structure across waves of data
- M*plus* software used with WLSMV estimator
  - First assessed plausibility of the one-factor CFA model
  - Then tested invariant factor loadings and item thresholds

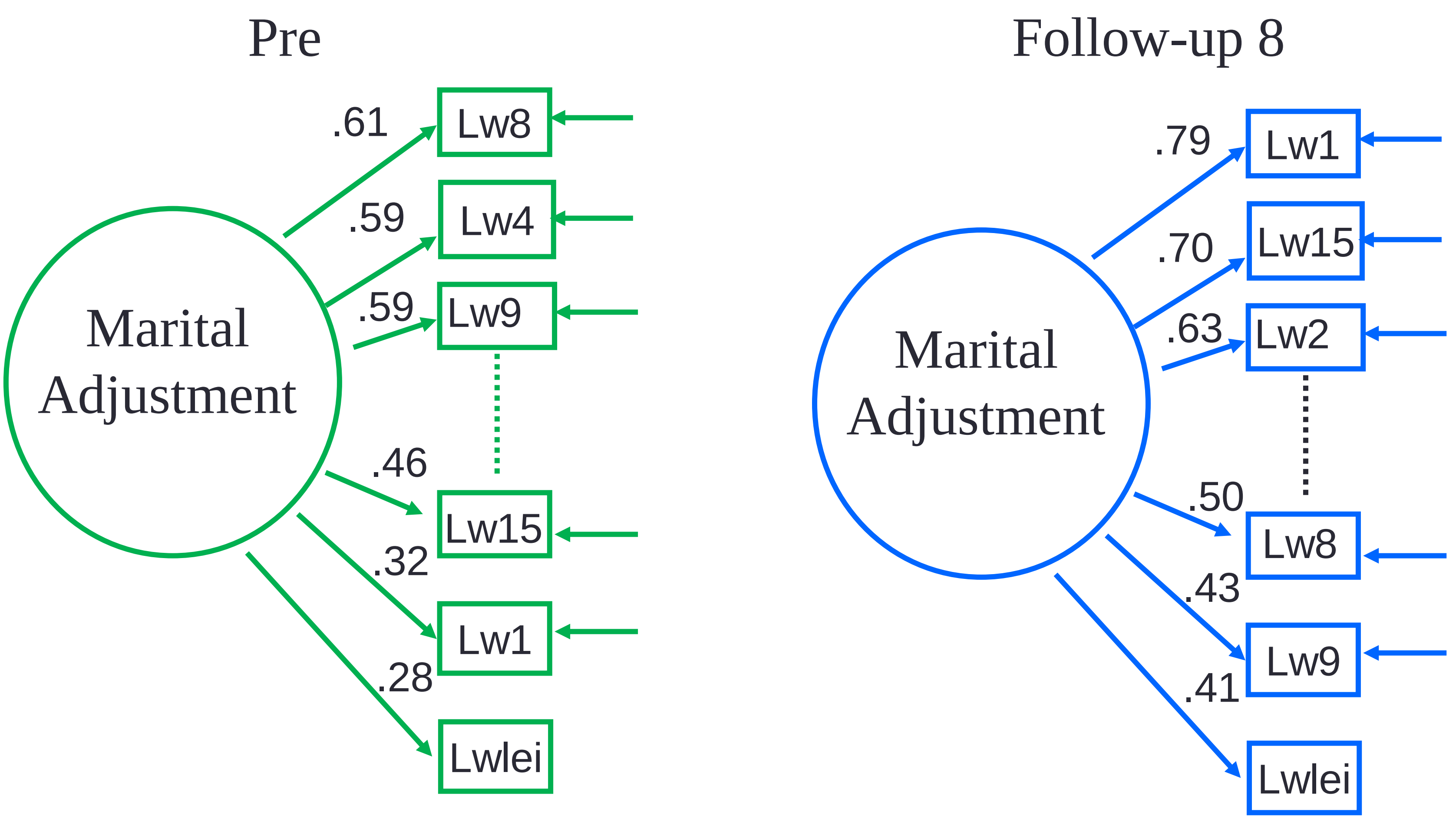
Multiple Groups CFA Results

- Two items, 14 and 16, were deleted due to low variance and contribution to lack of overall model fit, particularly in early waves
- The single-factor model fit the data fairly well across most waves, with some minor modification
  - Two item residual covariances were estimated across each wave, e.g., lw4 and lw6 (sex and affection)
- When testing invariance of adjacent waves, the factor structure, loadings, and thresholds were stable across most waves

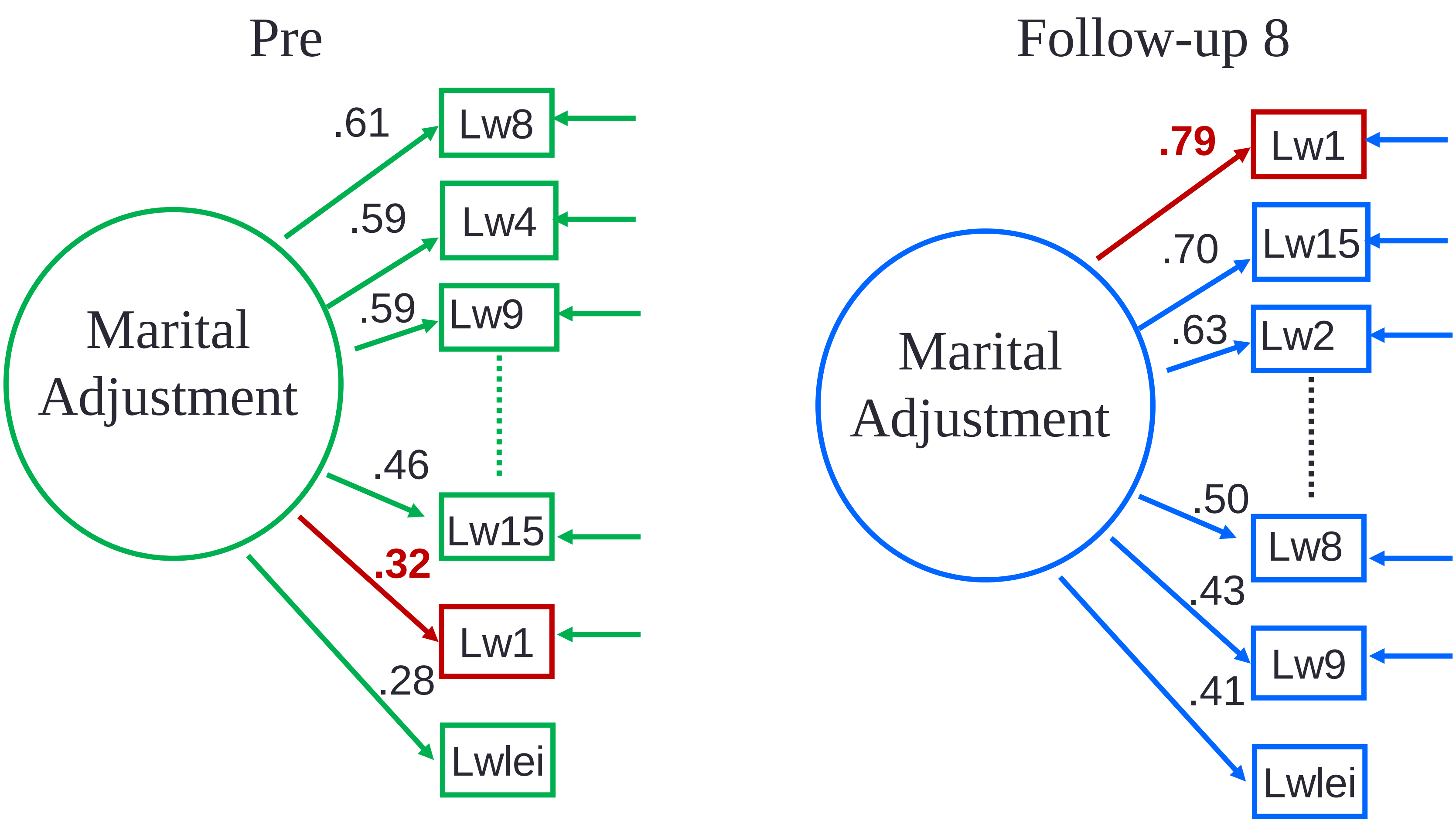
Multiple groups CFA results – cont’d

- However, some factor loadings began to “shift” across time
  - For example, loadings differed between the post assessment and first follow-up, between follow-up waves 2 and 3, and between pre/post and later waves
- Differences between non-adjacent waves indicated changes in magnitude **and** relative ordering of items
  - Several highest loading items during early phases became lowest later on, and vice versa
  - Patterns suggest construct definition changed across time

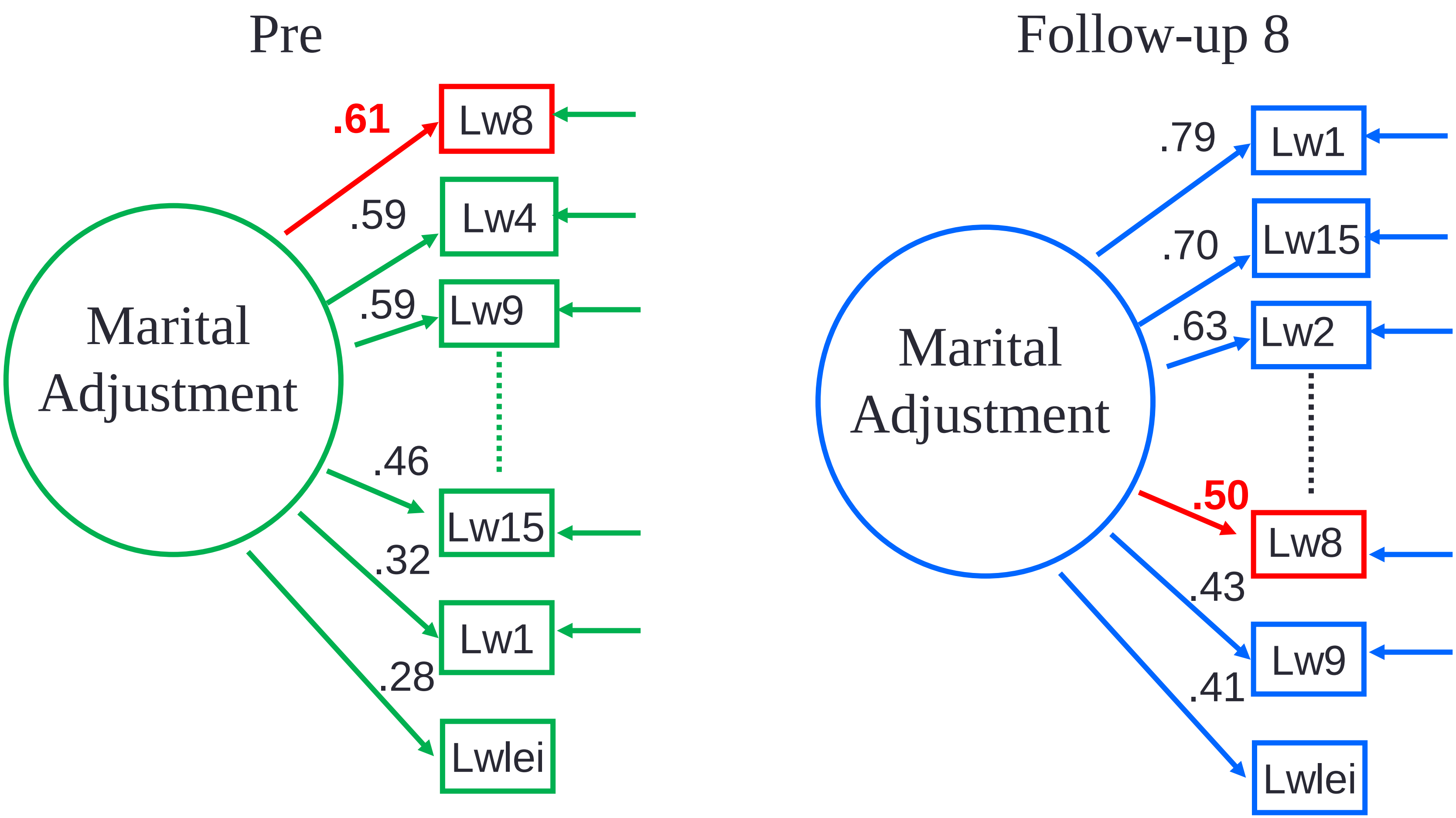
Loading Noninvariance Illustrated



Loading Noninvariance Illustrated



Loading Noninvariance Illustrated



Conclusions and Recommendations

- Evaluators interested in examining change across time need to ensure measurement equivalence prior to conducting tests of mean difference
- For standardized achievement tests used across different grade levels, this is usually accomplished by vertical equating
  - Though this approach has also been found inadequate as time intervals increase

Conclusions and recommendations – cont'd

- For measures of affective traits, evaluators should consult, *a priori*, theoretical and empirical evidence supporting stability of the trait
- Findings of nonequivalence "after the fact" leave few options
  - Delete nonequivalent items if possible
  - Conduct "think-aloud" protocols
  - Use the nonequivalence itself as something informative about the nature of changes across time

Take-home Message

- Documentation of prior reliability and validity does not ensure that scores/inferences from your sample are reliable/valid
  - Evaluators should conduct reliability and validity analyses for their sample
- Longitudinal research further requires evidence of reliable/valid scores for each wave of data
- Tests of mean differences cannot be trusted without evidence of equivalent measurement across time

Contact Information

- Antonio Olmos
  - polmos@du.edu

- Susan Hutchinson
  - susan.hutchinson@unco.edu