


Beyond Jacobson & Truax: Modern Methods for Estimating Clinically Significant Change



Antonio A. Morgan-López, Ph.D.
C-DIAS PSMG Virtual Grand Rounds
16 January 2024

Clinical Significance and *Individual-Level Outcomes*

- “.....clinically significant change as the extent to which therapy moves **someone** outside the range of the dysfunctional population or within the range of the functional population (Jacobson & Truax, 1991, p.12).”
- “...the behavior of the **target subject** is compared with that of his or her peers who have not been identified as problematic....behavior changes can be viewed as clinically important if the intervention has brought **the client's** performance within the range of socially acceptable levels (Kazdin, 1977, p.427).”

Clinical Significance: Reporting Requirements *in Parallel* to Effect Size

- “Although **effect sizes are useful ways of communicating the magnitude of a treatment effect**, they do not necessarily **communicate information about the clinical meaningfulness of an intervention**....authors are encouraged to use one of several approaches that have been recommended for capturing clinical significance, including (but not limited to) **the reliable change index (i.e., whether the amount of change displayed by a treated individual is large enough to be meaningful)**..[or] the extent to which dysfunctional **individuals** show movement into the functional distribution (La Greca, 2005, p.3)

Clinical Significance: Conflation with Effect Size

- “...Is the amount of change exhibited by an individual participant large enough to be considered meaningful (e.g., reliable change index; Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991), and are **treated individuals as a group** indistinguishable from normals with respect to the primary complaints following treatment (Kendall, 1999)

Clinical Significance: Conflation with Effect Size

Odgaard and Fowler (2010) in their recording of measures of clinical significance, effect size and confidence intervals from studies published in *Journal of Consulting and Clinical Psychology* from 2003-2008:

- “Finally, we recorded whether each article reported any measure of clinical significance.....(a) no clinical significance found, because comparison **groups** were equivalent on whatever metric used;.....(d) the comparison of **observed [effect sizes] with those previously published as clinically significant** (p.289).”

Clinical Significance v. Effect Size: Why the Distinction Matters



- Averaged treatment effects (e.g., raw/standardized mean differences in change) convey little-to-nothing about any specific individual (Jensen & Corralejo, 2017; Ogles et al., 2001)
- Possible to have large effect sizes on average and a non-trivial proportion of patients/participants who fail to improve or even get worse (Saavedra et al., 2021, 2022; Westen et al., 2004)

Clinical Significance I: Movement Below a “Normative Threshold” (Jacobson & Truax, 1991)

- Distinguishing between the clinical and non-clinical distributions on a scale score
- Requires an agreed upon community standard/grouping variable against which “clinical” and “non-clinical” groupings are defined (e.g., DSM diagnosis*)
 - Non-clinical examples:
 - Non-disordered external comparison samples (Kendall et al., 1999)
 - Patients that screened out of treatment RCTs (Saavedra et al., 2021, 2022)

*Criticisms of DSM dx notwithstanding (e.g., M-L et al., 2020, *JAD*; M-L et al., 2021, *JTS*; M-L et al., 2023, *IJMPPR*)

Clinical Significance I: Movement Below a “Normative Threshold” (Jacobson & Truax, 1991)

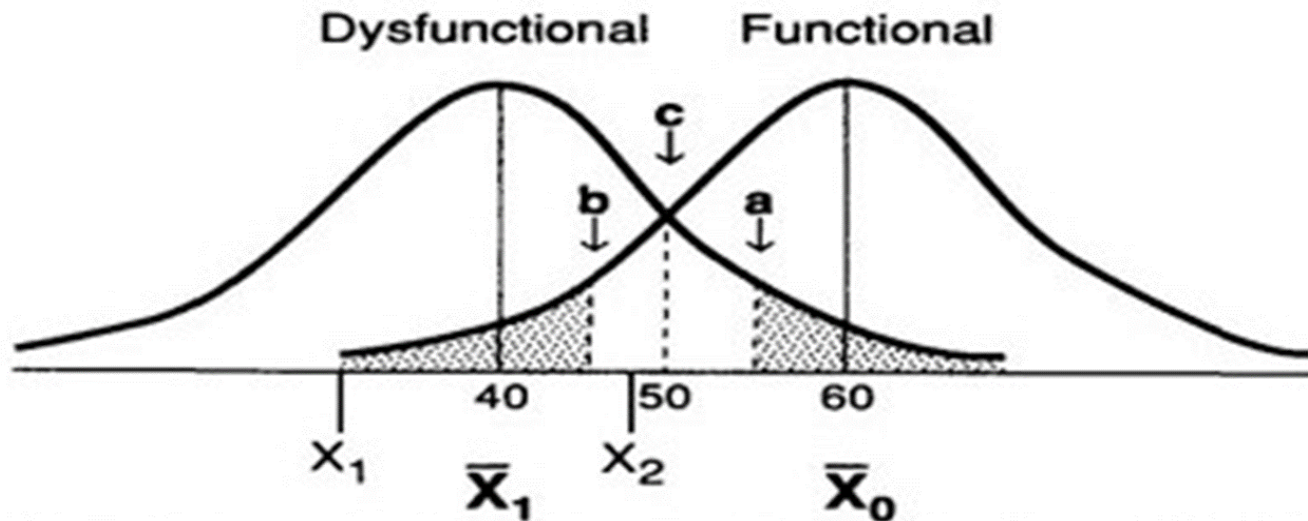


Figure 1. Pretest and posttest scores for a hypothetical subject (x) with reference to three suggested cutoff points for clinically significant change (a, b, c).

Clinical Significance I: “Normative Threshold” = Weighted Midpoint of Two Distributions

- $NT = \frac{\mu_{clinical}\sigma_{normative} + \mu_{normative}\sigma_{clinical}}{\sigma_{normative} + \sigma_{clinical}}$
- CSC is achieved if post-treatment scale score for patient $p < NT$

Clinical Significance II: Reliable Change Index (RCI; Jacobson & Truax, 1991; La Greca, 2005)

- “....the reliable change index (i.e., whether the amount of change displayed by a treated individual is large enough to be meaningful)....”

Clinical Significance II: Index RCI (Jacobson & Truax, 1991, *JCCP*)

$$RCI: \frac{d_i}{SEM_d}$$

- Assessment of whether an individual's change is significantly different from 0 (i.e., an *individual-level* significance test)
 - d_i = pre-post difference score for person i
 - SEM_d = standard error of measurement
 - Usually (but not always) grounded in internal consistency

RCI Inference Groupings

- Group individuals based on statistical significance ($p \leq .20$, Wise 2004) and direction of effect:
 - Statistically Significant Improvement (SSI)
 - Non-Significant Improvement (NSI)
 - Statistically Significant Deterioration (SSD)
 - Non-Significant Deterioration (NSD)

RCI: Continuing Popularity and Limitations

- Total citations of JT (91) ~ 13K
 - > 800 in 2023 alone
- Three specific limitations of the RCI
 - d_i based on two timepoints
 - d_i typically based on a total score psychometric model
 - SEM_d (i.e., reliability) is assumed to be universal across all persons and timepoints

RCI Limitation I: “What if I have multiple timepoints?”

- The numerator of the RCI estimate: “pre-post” difference scores
 - limited to an arbitrary 2nd timepoint versus using all timepoints

Multiple Timepoint RCI under MLM

- Speer and Greenbaum (1995; see also Lovaglio & Parabiaghi, 2014)
 - Proposed using MLM for the RCI
 - “Raw” random slope for patient i under an MLM/LGM = Empirical Bayes estimate of d_i
 - Not a lot of use in the literature (despite > 800 cites)
 - Typical use is with sum scores assuming perfect reliability
 - Fail to account for measurement bias and/or error

RCI Limitation II: Measurement Imprecision in d_i using Total Scores

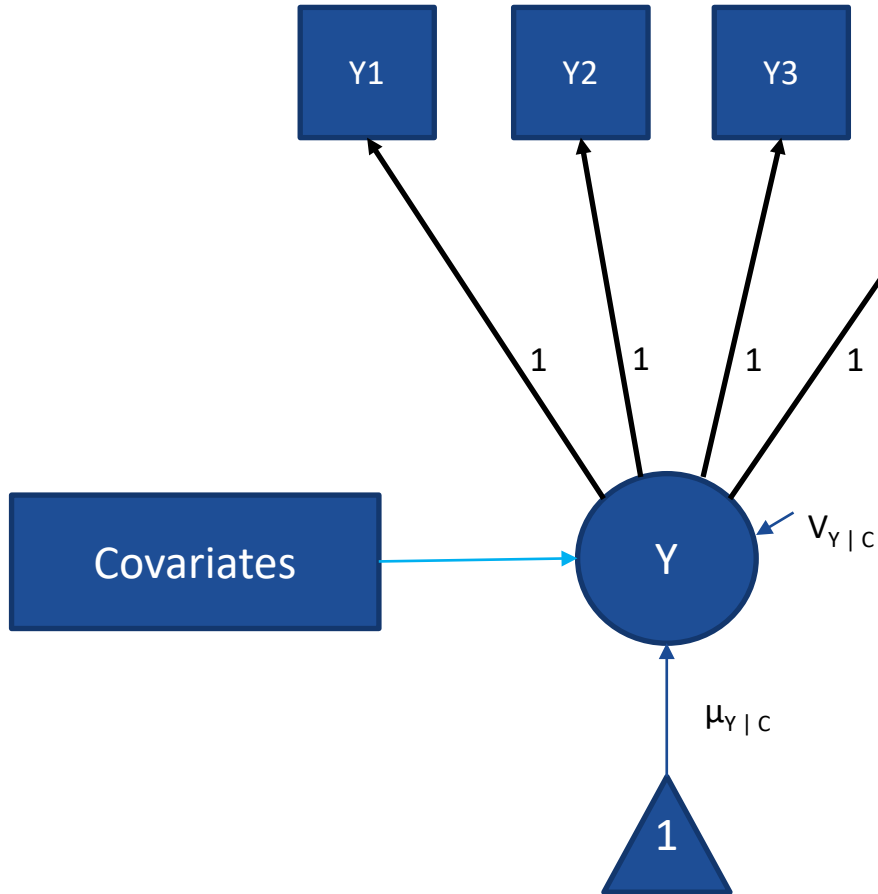
- Convenient to calculate (Curran et al., 2008)
- Seen largely as a “data management” problem moreso than an *untested* psychometric model (McNeish, 2022; McNeish & Wolf, 2020; Morgan-López et al., 2020, 2023; Saavedra et al., 2021, 2022)
- Underlying psychometric model will rarely-if-ever fit psychiatric outcome data (Andrich, 1978; He et al., 2014; McNeish & Wolf, 2020)

Total Score Model Assumptions

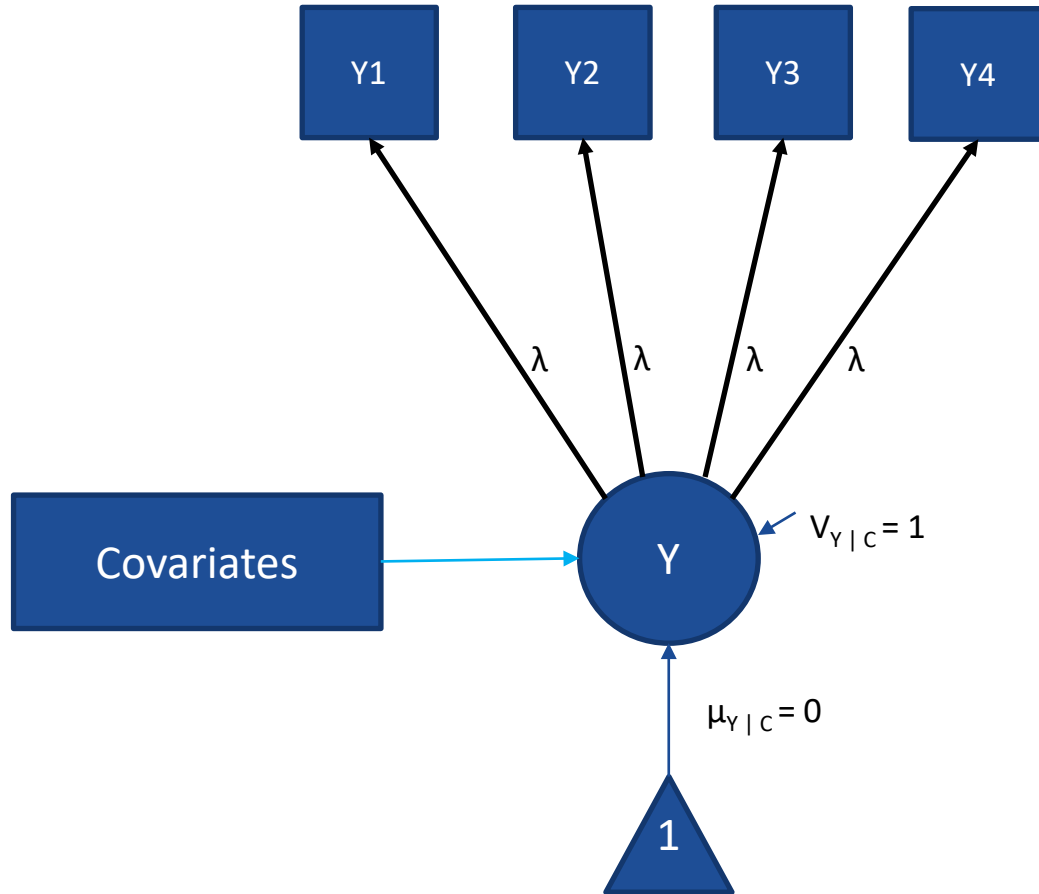
- The relative weight of each item/symptom must be equal
 - In FA/IRT, **equal factor loadings across items/symptoms**
- *Neither* the relative weight nor prevalence of each symptom should differ across populations, time, reporters, etc., (above-and-beyond “true” differences/change in the construct)
 - **Measurement Invariance**
- **Testable psychometric model** under factor analysis/item response theory (FA/IRT)

(McNeish & Wolf, 2020, *BRM*; Morgan-López et al., 2022, *JTS*)

Unstandardized “Total Score” Model in Generalized Factor Analysis (GFA)



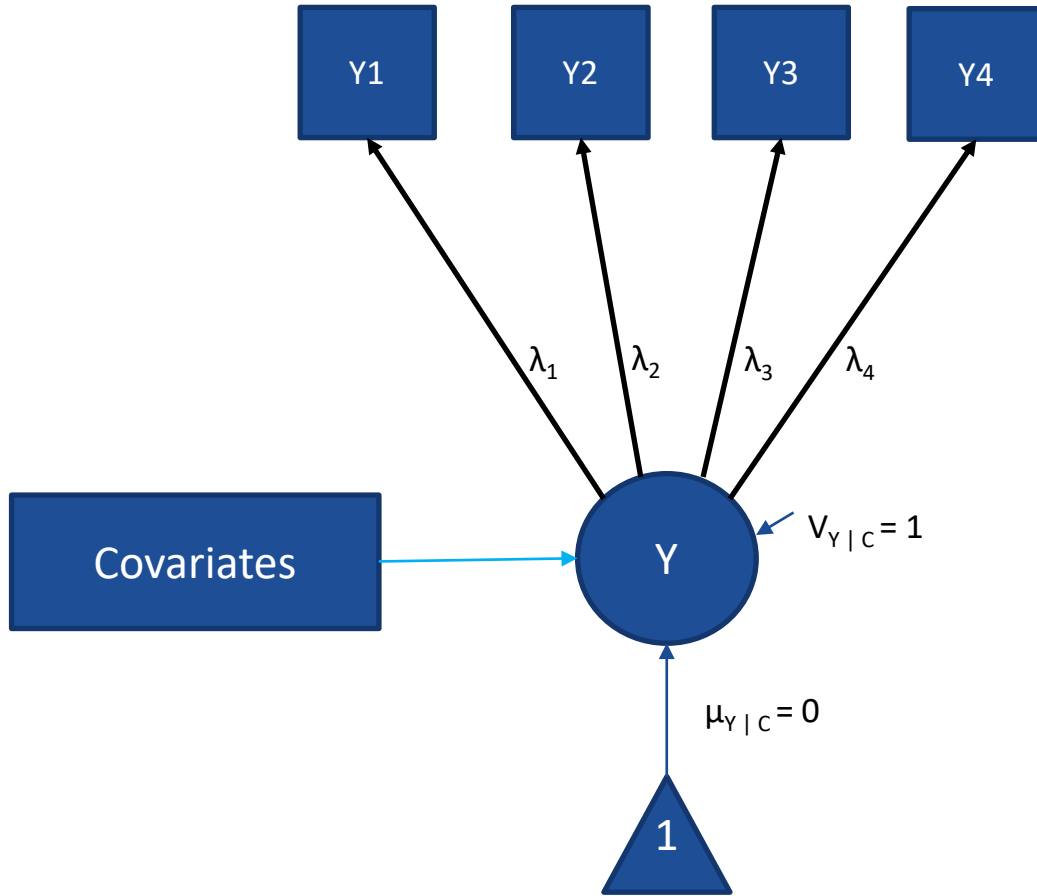
Standardized “Total Score” Model in Generalized Factor Analysis (GFA)



Relaxing the “Equal Weights” Assumption: (G)FA

- More “typical” factor analysis structure
 - The relative weights (i.e., factor loadings) can vary across items/symptoms
- Still assumes measurement invariance across populations, time, reporters, etc., (above-and-beyond “true” differences/change in the construct)

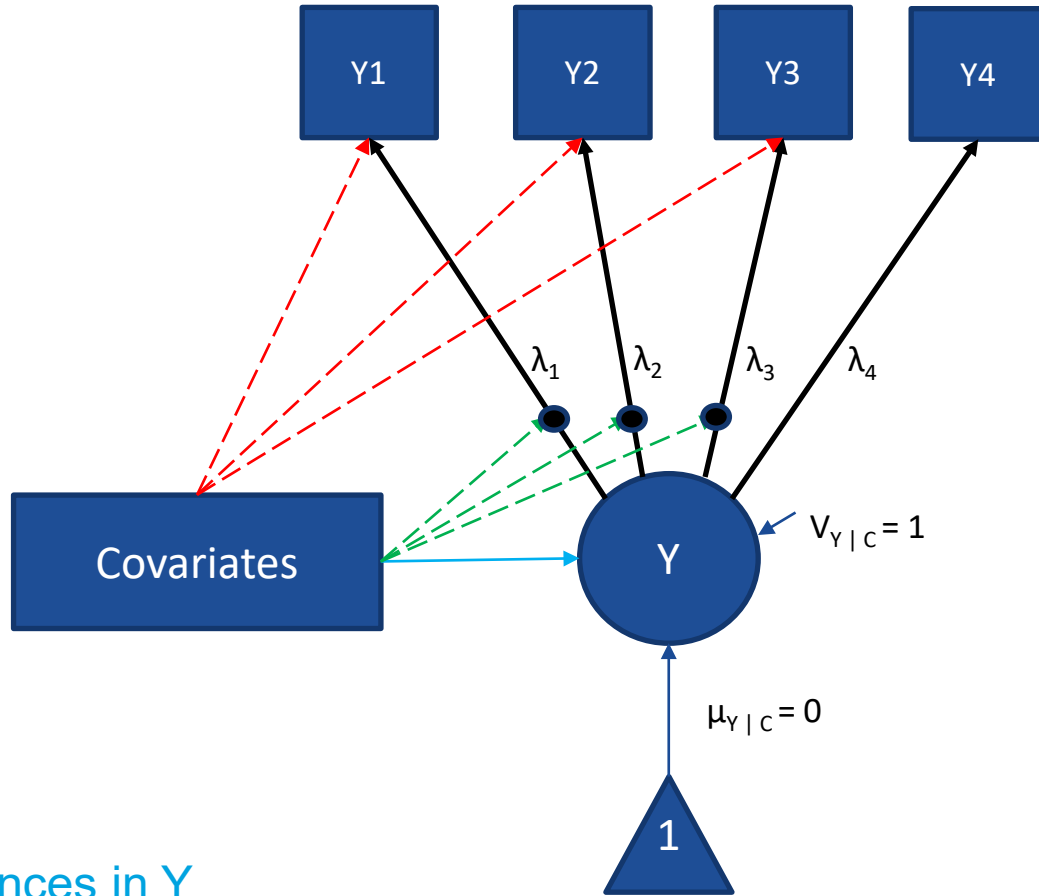
Conventional GFA



Measurement Non-Invariance in (G)FA

- Item thresholds and factor loadings can vary across items/symptoms
- Different thresholds and/or factor loadings may be necessary for some items/symptoms across:
 - Time, Populations, Reporters, *Studies*
- Old-School Methods
 - Multiple-Group GFA (& IRT)
 - Multiple Indicator, Multiple Cause (MIMIC) models
- New-School Methods
 - Moderated nonlinear factor analysis

Advanced Scale Scoring under Moderated Nonlinear Factor Analysis (MNLFA; Bauer, 2017)

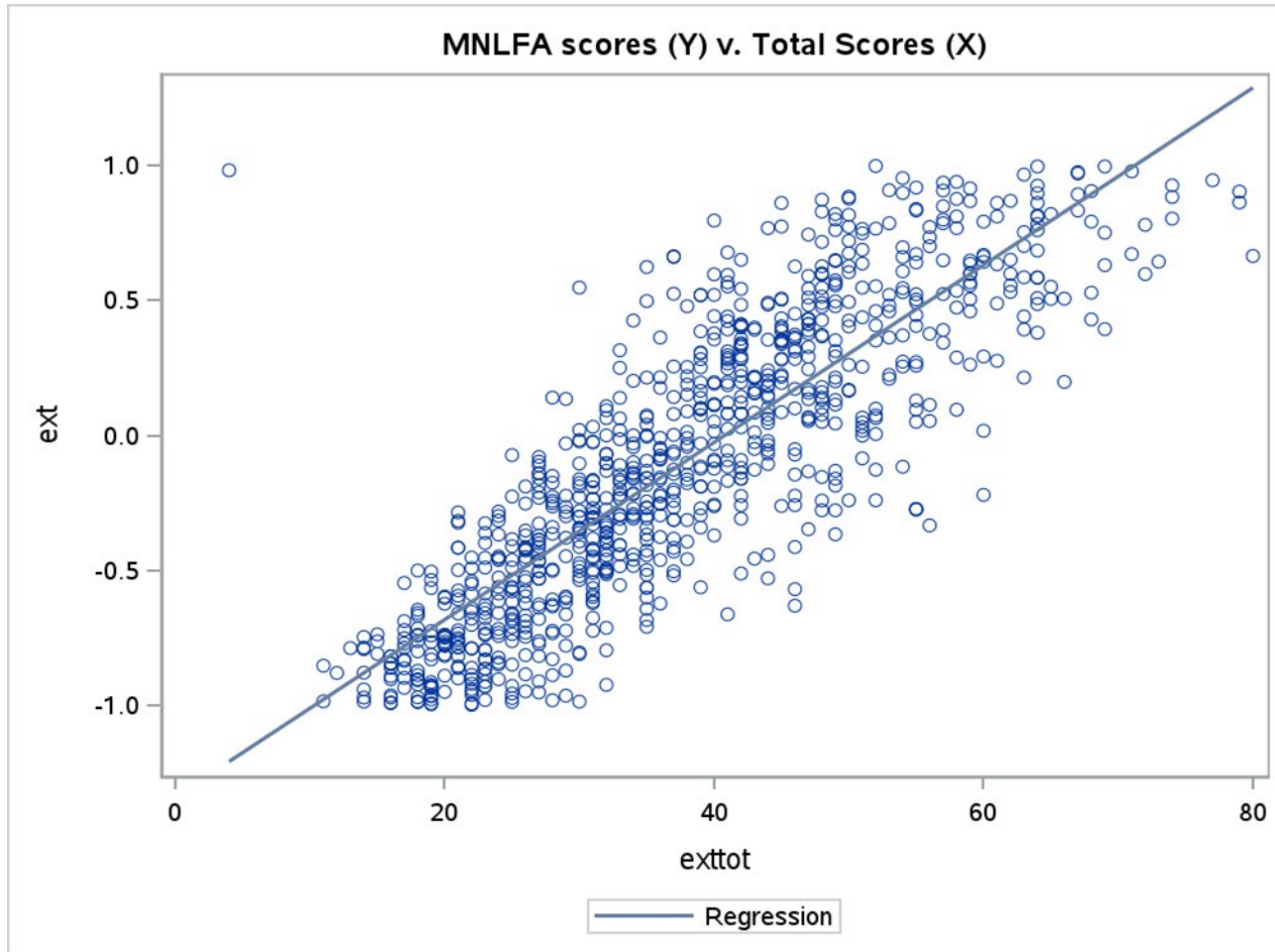


“True” Differences in Y

Measurement Differences in Item “Severity”

Factor Loading/“Discrimination” Differences

“If sum/total scores and IRT/FA scores are so highly correlated, why can’t I just use total scores?” A cautionary graphic (M-L et al., R&R-a, *Behavior Therapy*)



$$r_{\text{total, MNLFA}} = \underline{\underline{.94}}$$

“If sum/total scores and IRT/FA scores are so highly correlated, why can’t I just use total scores?” A cautionary graphic (M-L et al., R&R-a, *Psychiatric Services*)



$$r_{\text{total, MNLFA}} = \underline{\underline{.95}}$$

Sum/Total Scores versus MNLFA Scores: Real Implications for Inferences (McNeish & Wolf, 2020; McNeish, 2022)

- If the psychometric model underlying sum/total scores does not fit the data *but they are used anyway*
 - Effect size distortion (Type I or Type II errors)
 - General between- and within-group differences
 - **Clinical decision-making errors at the individual-level**
 - “Over” and “Under” Diagnosis
 - Premature Treatment Discharge
 - Particular implications for distortion of outcomes among minoritized populations

RCI Limitation III: Imprecision with SEM_d

- SEM_d for conventional RCI (under classical test theory) =
 - $S (\sqrt{1 - r})$
 - S = standard deviation of the scale scores, r = reliability
 - Assumed constant for everyone at every timepoint
- SEM_{di} under NLFA/IRT
 - $1 / (\sqrt{TIF_i})$
 - Test Information Function (TIF) value for observation i (SEM_{di} in factor analysis/IRT scale score output dataset)

Limitations of the Original RCI

- d_i is based on two timepoint difference score
 - Individual MLM/LGM slopes (Speer & Greenbaum, 1995; Lovaglio & Parabiaghi, 2014)
- d_i is biased under CTT; SEM_d biased and presumed equal across participants and time
 - Estimate scores and SEMs under advanced factor analysis or IRT (Brouwer et al., 2013; Saavedra et al., 2021)
- Can lead to 25-55% RCI misclassification
 - Jabrayilov et al., 2016; M-L et al., 2022; M-L et al., post-R&R; Saavedra et al., 2021, 2022

Multiple Timepoint, Multiple Error Source RCI Model (M-L et al., 2022a, R&R; Saavedra et al., 2022)

- Based on (a modification of) the measurement error-corrected multilevel model (Wang et al., 2019; Zhang et al., 2021)
- Estimate RCIs accounting for measurement error AND measurement bias:
 - Variation in item parameters (i.e., MNI/DIF) across time, populations, etc. (prior to MLM/LGM)
 - Accounting for multiple sources of error:
 - Prediction error (“Conventional” MLM/LGM Level-1 residual)
 - Uncertainty in the score estimate for θ across multiple timepoints and patients

Unconditional 2-Level Longitudinal MLM

Level-1 Model: $Y_{it} = \beta_{0i} + \beta_{1i} (\text{Time}_{it}) + r_{it}$

Level-2 Model: $\beta_{0i} = \beta_{00} + g_{0i}$
 $\beta_{1i} = \beta_{10} + g_{1i}$

MEC-MLM (adapted for the RCI) (Diakow, 2013; Wang et al., 2019)

$$\text{Level-1: } Y_{it} = \hat{\theta}_{it} + s_{it}$$

$$\text{Level-2: } \hat{\theta}_{it} = \beta_{0i} + \beta_{1i} (\text{Time}_{it}) + r_{it}$$

$$\text{Level-3: } \beta_{0i} = g_{0i}$$

$$\beta_{1i} = g_{1i}$$

r_{it} = deviation of Y_{it} from the predicted trajectory for person I

s_{it} = Drawn from **fixed** person- and time-specific SEM²(SE² from MNLFA scale scores. Similar to reading in fixed variances in meta-analysis; Sheu & Suzuki, 2001)

(M-L et al., 2022, *IJMPR*; M-L et al., R&R, *BT*; Saavedra et al., 2022, *BT*)

MEC-MLM (adapted for the RCI) (Diakow, 2013; Wang et al., 2019)

$$\text{Level-1: } Y_{it} = \hat{\theta}_{it} + s_{it}$$

$$\text{Level-2: } \hat{\theta}_{it} = \beta_{0i} + \beta_{1i} (\text{Time}_{it}) + r_{it}$$

$$\text{Level-3: } \beta_{0i} = g_{0i}$$

$$\beta_{1i} = g_{1i}$$

g_{0i} = “Raw” intercept for person i

g_{1i} = “Raw” slope for person i - $g_{1i} / SE_{g_{1i}}$ is the RC estimate

- *No fixed effects in this model(!!)

(M-L et al., 2022, *IJMPR*; M-L et al., R&R, *BT Saavedra et al.*, 2022, *Behavior Therapy*)

Advanced RCI Selective Prevention Example: The Coping Power Intervention (M-L et al., R&R, *BT*)

- “Standard” CP delivered in group format (GCP; Lochman & Wells, 2002a/b; Lochman et al., 2013)
- Other CP formats:
 - CP with Mindfulness (Boxmeyer et al., 2021; Miller et al., 2020)
 - Internet-delivered CP (Lochman et al., 2017)
- Individual CP (ICP; Lochman et al., 2015, 2019)
 - Developed to counteract concerns regarding potential iatrogenic peer contagion effects in GCP
 - Shown comparative efficacy versus GCP on average

Child Externalizing RCTs: Overreliance on Comparisons of Group-Averaged Trajectories?

Dodge, Dishion & Lansford (2006) quote:

- “Reporting only the **average effects masks variability in responses to an intervention.** Some interventions, especially those that aggregate [conduct disordered] youth, might result in average improvement across youth, **but serious deterioration for a sub-group of youth.** This possibility cannot be evaluated unless **individual responses are summarized in scientific reports** (p.14).”
- Summaries of individual responses = Clinically Significant Change (CSC)

Purpose of the Current Study

- Comparison of RCI classification %s on broadband externalizing between ICP and GCP (Lochman et al., 2015, 2019)
- Part of an 11-study Integrative Data Analysis (IDA) examining CP's distal effects on reductions in risk for suicidality and completed suicides
 - (McDaniel et al., 2022, *PS*; Morgan-López et al., 2022, *CCT*; Saavedra et al., 2024, *D&P*)

Participants

- Youth from 20 schools identified as at-risk for aggressive behavior in 4th grade (> 75thile on ABS)
 - Post-baseline assessments in summer of 5th grade and roughly yearly thereafter through 11th grade (Wave 8)
- 360 parent-child pairs
- School-level randomization
 - 10 schools randomized to ICP, 10 to GCP (6 youth per group within schools)
- Equal number of sessions (ICP = 28.96, GCP = 28.54; out of 32)

Measures

- Demographics (serve as a) predictors of MNI/DIF and b) auxiliary variables for multiple imputation)
 - Baseline age, Gender, Race/Ethnicity, SES
- Intervention Condition (ICP = 1, GCP = 0)
- Teacher-Reported BASC Externalizing
 - Estimated under bifactor MNLFA (Eid et al., 2016; Hussong et al., 2020)
 - Item parameters reported in Saavedra et al., (2024)

Baseline Descriptives

Table 1
Descriptives at Baseline

Variable	GCP		ICP		<i>p</i> -value
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age	10.2	0.51	10.24	0.49	0.43
Gender	0.69		0.63		0.23
Race/Ethnicity					0.48
	Black	82.4%	76.7%		
	White	14.1%	18.7%		
	Other	3.5%	4.7%		
Family Income					0.98
	None	4.9%	4.0%		
	<\$15,000	24.7%	24.5%		
	\$15,000-\$29,999	33.3%	31.8%		
	\$30,000-\$49,999	21.8%	23.2%		
	>\$50,000	15.5%	16.6%		

Notes: GCP = Group Coping Power. ICP = Individual Coping Power.

Analysis Plan

- BASC Externalizing Scale Score Estimation
 - Bifactor MNLFA model (Saavedra et al., 2024, *D&P*)
 - Alternative model: single factor NLFA with equality constraints across all factor loadings (which fit the data poorly....)
 - Tests the formal psychometric model underlying total/mean scores (McNeish & Wolf, 2020; McNeish, 2022; M-L et al., 2022c, 2023)
- Multilevel MI (Keller & Enders, 2021)
 - 50 MI datasets
 - MAR versus “R x T” NMAR
 - Graphical evidence suggesting youth who were a) in GCP and b) were lost to follow-up earlier were headed down worse externalizing trajectories

Final BASC Externalizing Item Parameters

Specific factor/item	Threshold (τ; 0 to 1)	Threshold (τ; 1 to 2)	Threshold (τ; 2 to 3)	Gender	Race	Age	Income	Repeat grade	T3	T4	T5	T6	T7	T8
Conduct														
Has to stay after school for punishment	0.61	3.18	5.42						1.07					
Steals/steals at school	1.80	3.89	5.37											
Cheats in school	0.37	2.97	4.79											
Uses foul language	-0.06	3.02	5.39								0.68		1.14	
Shows a lack of concern for others' feelings	-1.88	1.40	3.32											
Skips classes at school	3.65	5.65	7.23											
Complains about police or other law enforcement officers	2.98	5.05	6.64											
Is truant	1.53	3.21	4.60											
Has been suspended from school	0.44	3.33	5.56	0.59					1.30					
Has friends who are in trouble	-2.08	0.90	2.95		0.34							-1.30	-0.89	
Aggression														
Argues when denied own way	-3.61	-0.38	1.56											
Threatens to hurt others	-1.21	2.82	5.39											
Blames others	-3.88	-0.57	1.56									-1.56		
Bullies others	-1.95	1.71	4.16											
Breaks other children's things	0.54	3.32	5.13											
Talks back to teachers	-2.79	0.60	2.53											1.26
Orders others around	-1.78	1.14	3.53	-1.47			0.08							
Is critical of others	-2.38	1.06	3.32	-0.81										
Calls other adolescents names	-3.84	0.37	3.26											
Shows off	-1.87	0.68	2.83											
Teases others	-4.27	0.45	3.46											
Complains about rules	-2.03	1.30	3.68									1.61	1.45	1.59
Hits other adolescents	-1.20	2.70	5.28											
Is a "sore loser"	-1.33	1.63	3.77											
Hyperactivity														
Rushes through assigned work	-5.26	0.33	4.05											
Bothers other children when they are working	-3.18	0.51	3.07	0.49										
Talks too loud	-2.13	0.38	2.33	-0.85	0.57									1.14
Seeks attention while doing schoolwork	-2.28	0.57	2.54											
Taps foot or pencil	-0.66	1.53	3.20	0.68										
Acts without thinking	-3.89	-0.09	2.31					0.53						
Calls out in class	-2.95	0.31	2.58									0.87		
Interrupts others when they are speaking	-3.23	0.85	3.45											
Makes loud noises when playing	-0.19	1.68	3.33											
Hurries through assignments	-4.17	0.26	3.15											
Acts silly	-2.77	0.03	1.81											
Is overly active	-1.23	1.03	2.73	0.43										
Cannot wait to take turn	-1.08	1.94	3.98											

Note. BASC = Behavior Assessment System for Children; MNLFA = moderated nonlinear factor analysis; DIF = differential item functioning.

Table 3. Final BASC MNLFA parameters: loading DIF

Specific factor/item	General factor loading (λ)	Specific factor loading (λ)	Gender	Race	Age	Income	Repeat grade	T3	T4	T5	T6	T7	T8
Conduct													
Has to stay after school for punishment	1.11	0.58						0.63					
Steals/steals at school	1.09	0.77											
Cheats in school	1.11	1.03											
Uses foul language	1.47	1.62											
Shows a lack of concern for others' feelings	1.71	0.69								0.58			
Skips classes at school	0.00	2.13											
Complains about police or other law enforcement officers	1.07	1.33											
Is truant	0.52	1.16											
Has been suspended from school	1.29	1.58											
Has friends who are in trouble	1.18	0.80											
Aggression													
Argues when denied own way	2.18	0.53											
Threatens to hurt others	2.14	2.19											
Blames others	2.21	0.85											
Bullies others	2.48	2.09											
Breaks other children's things	1.66	0.89											
Talks back to teachers	2.25	0.88											
Orders others around	2.01	0.96			0.56								
Is critical of others	2.22	0.75											
Calls other adolescents names	2.43	1.68											
Shows off	1.95	0.00											
Teases others	2.85	1.77											
Complains about rules	1.94	0.51											
Hits other adolescents	1.75	1.66											
Is a "sore loser"	1.78	0.56											
Hyperactivity													
Rushes through assigned work	2.21	3.78											
Bothers other children when they are working	2.46	0.50											
Talks too loud	2.18	0.00											
Seeks attention while doing schoolwork	2.19	0.00											
Taps foot or pencil	1.13	0.00											
Acts without thinking	2.35	0.62											
Calls out in class	2.74	0.00											
Interrupts others when they are speaking	2.90	0.00											
Makes loud noises when playing	1.62	0.00											
Hurries through assignments	1.69	2.62											
Acts silly	1.34	0.00											
Is overly active	2.00	0.00											
Cannot wait to take turn	2.11	0.00											

Note. BASC = Behavior Assessment System for Children; MNLFA = moderated nonlinear factor analysis; DIF = differential item functioning.

Dataset (note that 'ext' and 'ext_se' were output from Mplus measurement models)

SAS Enterprise Guide

File Edit View Tasks Favorites Program Tools Help [Icons] Process Flow

Project Tree - x BASC Externalizing CSC-RCI PART II (Cheung-McN, MAR multiple imputation) -

Program Log Output Data (2)

WARN -

Filter and Sort Query Builder Where Data Describe Graph Analyze Export Send To

	id_u	IGCP	TIMEPT	ext	ext_se	exttot	exttot_se
1	1	0	0	1.093	0.212	65	3.9903376048
2	1	0	2	0.492	0.217	44	3.9903376048
3	1	0	3	-0.119	0.221	42	3.9903376048
4	2	0	0	0.567	0.23	41	3.9903376048
5	2	0	2	-0.241	0.213	28	3.9903376048
6	2	0	3	-0.896	0.227	17	3.9903376048
7	2	0	4	0.405	0.222	45	3.9903376048
8	3	0	0	-0.25	0.235	28	3.9903376048
9	3	0	2	0.595	0.225	41	3.9903376048
10	3	0	3	0.141	0.237	28	3.9903376048
11	4	0	0	-0.846	0.255	24	3.9903376048
12	4	0	2	-0.76	0.221	27	3.9903376048
13	4	0	3	-0.613	0.22	31	3.9903376048

“Easier” RCI Estimation under MEC-MLM (Cheung, 2008; McNeish, 2016)

- Estimate a “weight” based on the inverse of the SE^2 of the (MNLFA) score from your Mplus output dataset
- Estimate an RCI model with no measurement error
 - Save growth parameter variances/covariances and residual variances
- Fit MEC-MLM
 - Use GPV/Cs and RVs as start values
 - Constrain the level-1 measurement error variance to 1
 - Use the $1 / SE^2$ as a weight

“Easier” RCI Estimation under MEC-MLM (Cheung, 2008; McNeish, 2016)

```
libname m 'BT Supplement';  
data rci_long; set m.rci_long;  
  
obs=_n_;  
weight=1/(y_se**2);  
  
title "RCI - Cheung/McNeish specification";  
proc mixed data=rci_long covtest method=reml noclprint;  
class i obs;  
model y= / noint ddfm=kenwardroger;  
random int time/subject=i type=un;  
random int/subject=obs(i) type=vc solution;  
parms (.09) (.04) (.03) (1.04) (1) / hold=5;  
weight weight;  
ods output SolutionR=random(rename=(stderrpred=StdErr));  
run;quit;
```

RCI Output Dataset from MEC-MLM (M-L et al., 2022; Saavedra et al., 2022)

se Guide

view Tasks Favorites Program Tools Help Process Flow

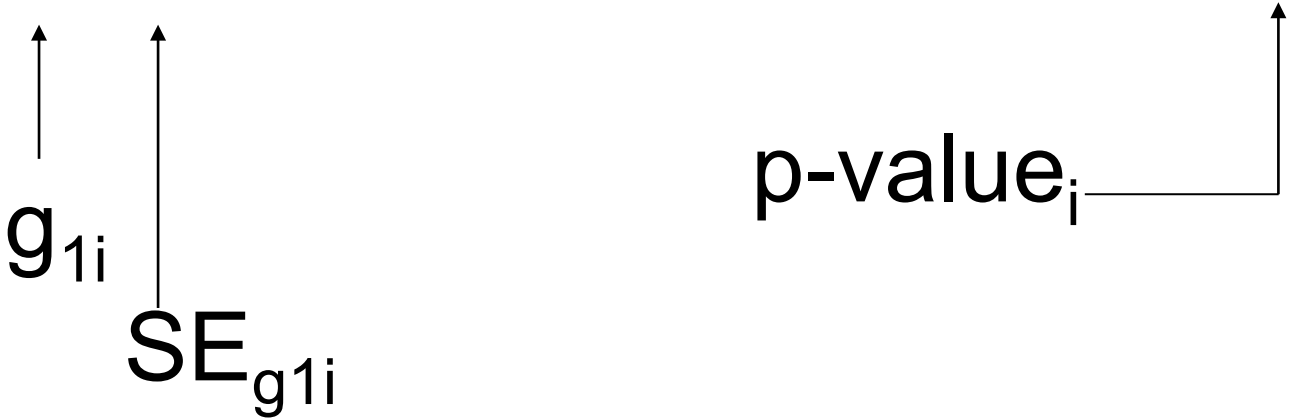
MAR outcomes analysis

Program Log Output Data (7) Results

R_IMP

Filter and Sort Query Builder Where Data Describe Graph Analyze Export Send To

	id_u	NImpute	Parm	Estimate	StdErr	LCLMean	UCLMean	DF	Min	Max	Theta0	tValue	Probt
1	1	20	timept	-0.120520	0.129987	-0.37972	0.1387	70.744	-0.261946	0.038235	0	-0.93	0.3570
2	2	20	timept	-0.109036	0.109424	-0.32489	0.1068	188.78	-0.203078	0.034854	0	-1.00	0.3203
3	3	20	timept	-0.037935	0.105172	-0.24502	0.1691	264.25	-0.110983	0.039545	0	-0.36	0.7186
4	4	20	timept	-0.146048	0.113005	-0.36940	0.0773	144.95	-0.310011	-0.036600	0	-1.29	0.1983
5	5	20	timept	-0.089852	0.135877	-0.36152	0.1818	61.381	-0.327822	0.069398	0	-0.66	0.5109
6	6	20	timept	-0.095308	0.115993	-0.32493	0.1343	121.68	-0.215853	0.029876	0	-0.82	0.4129
7	7	20	timept	-0.145651	0.113934	-0.37087	0.0796	142.74	-0.247358	-0.009879	0	-1.28	0.2032
8	8	20	timept	-0.000551	0.114105	-0.22627	0.2252	131.51	-0.161367	0.091868	0	-0.00	0.9962
9	9	20	timept	-0.071451	0.108408	-0.28525	0.1423	195.63	-0.195370	0.051450	0	-0.66	0.5106
10	10	20	timept	-0.132152	0.116537	-0.36292	0.0986	118.47	-0.241282	0.014033	0	-1.13	0.2591



Individual-Level Longitudinal Cohen's d (Feingold, 2019)

$$\frac{g_{1i} \times (\text{number of timepoints} - 1)}{SD_Y \text{ at Baseline}}$$

RCI Results: Agreement Across Scoring Methods

- Bifactor MNLFA models versus Total Scores
 - Only 43.6% agreement in RCI inferences
 - Test of Symmetry, $\chi^2(15) = 31.23$, $p=.008$
 - Weighted Kappa: .009 (95% CI: -.082, .100)
- Bifactor MNLFA models versus Bifactor NLFA (assuming MI)
 - 82.5% agreement in RCI inferences (still 1 out of 5 kids misclassified)
 - Test of Symmetry, $\chi^2(15) = 7.99$, $p=.92$
 - Weighted Kappa: .829 (95% CI: .788, .869)

RCI Results: ICP/GCP Differences by Scoring Method

Table 5

RCI Classifications by Scoring Method: Bifactor MNLFA Scores, Bifactor Scores without DIF, and Total Scores

RCI Inference (at $p < .20$)	Total Scores		Bifactor Scores without DIF		Bifactor MNLFA Scores (from Table 4)	
	ICP (n = 180)	GCP (n = 180)	ICP (n = 180)	GCP (n = 180)	ICP (n = 180)	GCP (n = 180)
Deterioration, $d > .5$	0	0	1.1	1.1	1.1	1.7
Deterioration, $.2 < d < .5$	0.6	0	0.6	4.5	0.6	5.7
Deterioration, $d < .2$	1.0	2.8	2.8	9.0	3.3	10.6
Improvement, $d < .2$	4.0	13.0	5.7	8.5	6.7	14.4
Improvement, $.2 < d < .5$	24.9	32.8	15.3	24.3	15.0	22.2
Improvement, $d > .5$	69.5	51.4	74.6	52.5	73.3	45.6

Notes: Measurement Error-Corrected Multilevel Models fit under multiple imputation for non-ignorable missingness (Demirtas & Schafer, 2003). RCI = Reliable Change Index. GCP = Group Coping Power. ICP = Individual Coping Power. DIF = Differential Item Functioning. MNLFA = moderated nonlinear factor analysis.

Random effects multinomial logit models (school-level RIs):

From RCI/NMAR Bifactor MNLFA model: $\chi^2(5) = 17.01$, $p = .005$, $\phi = .22$

Conclusion from M-L et al., (R&R, *BT*)

- There have been long-standing calls for:
 - Examination of individual-level improvement/deterioration in RCTs, particularly for youth who are at early risk for conduct problems
 - Improvements in accuracy/precision/flexibility in estimating CSC more broadly
- In this reexamination of the IGCP trial using “modernized” CSC methods:
 - The proportions of youth who saw meaningful reductions in externalizing were at least as great for ICP compared to GCP
 - ICP > GCP under the NMAR model
 - 20-30% of youth failed to improve/got worse (not an atypical finding)
 - Significantly higher % in GCP, underscoring concerns regarding iatrogenic effects in some youth

Advanced RCI Treatment Example: COPE versus Relapse Prevention (Saavedra et al., 2022, BT)

- Assessment of whether *each individual patient's change* in PTSD severity is a) significantly different from 0 or b) below a normative threshold
 - All-VA sample from RCT comparing COPE and RP (Back et al., 2018; n = 81)
 - Normative comparison sample that were screened out of the RCT because they did not meet for PTSD (“normative” n = 48)
- Are RCI/NT inferences (e.g., sig improvement, sig deterioration) affected by:
 - Use of CAPS-IV total scores versus (MNL)FA scores
 - MNI/DIF across multiple factors (including R/E)

Table 2
Final CAPS-IV MNLFA Item Parameters

PTSD Symptom	Threshold/ Difficulty	Loading/ Discrimination	Age Threshold MNI/DIF	Black Threshold MNI/DIF	Married Threshold MNI/DIF	Normative Threshold MNI/DIF	Pre- Treatment Rx Threshold MNI/DIF	Week 6 Threshold MNI/DIF	Week 6 Loading MNI/DIF
Intrusive Recollections	-3.12	1.86						-1.66	-1.24
Nightmares	-1.79	0.93	0.04						
Flashbacks	0.17	1.21							
Psychological Cues	-2.38	1.38							
Physiological Cues	-2.19	1.33					0.80		
Thought Avoidance	-2.44	1.45							
Activity Avoidance	-0.96	1.01							
Inability to Recall	0.70	0.47		-1.23					
Diminished interest	-1.73	1.16			1.09				
Detachment	-2.15	1.21	-0.04			-1.09			
Restricted Affect	-2.28	1.19							
Foreshortened Future	0.44	0.77							
Sleep	-2.32	0.70							
Irritability	-1.51	0.83	-0.05						
Concentration Probs	-1.75	1.02		-1.10			-0.77		
Hypervigilance	-2.16	0.90							
Startle	-0.79	0.65							

Note: For MNI/DIF parameter estimates, estimates that were significant at $p < .05$ are shown and included in the final scoring model.

Saavedra, M-L, Back et al., (2022, *BT*)

RCI Inference (at $p < .20$)	CAPS MNLFA		CAPS Symptom Counts	
	RP (n = 27)	COPE (n = 52)	RP (n = 27)	COPE (n = 52)
Significant Deterioration	0.0	0.0	0.0	0.0
Non-significant Deterioration	37.0	17.3	3.7	1.9
Non-significant Improvement	22.2	7.7	7.4	9.3
Significant Improvement	40.7	75.0	88.9	88.9

Saavedra, M-L, Back et al., (2022, *BT*)

- Using total symptom counts, COPE and RP had an equal impact on a) the % of patients with significant improvement and b) reductions in the number of symptoms
- Using MNLFA scores,
 - COPE reduced a set of symptoms that were more “difficult” to treat than RP
 - Higher %s of patients with SSI when considering the relative weighting of symptoms

Overall Conclusion

- Consequences for inference if measurement is treated simply as “data management”
 - Particularly critical for estimation of clinical significance trajectories
- Novel methods for RCI estimation
 - Advanced scale score estimation under MNLFA (e.g., in Mplus or R)
 - Adaptation of MEC-MLM (in SAS, SPSS, R/’lme4’, Mplus*)
- An “old”/new way of characterizing RCT results (Kazdin, 1977; Jacobson & Truax, 1991)
 - Differences in the % of patients who improve, fail to improve, get worse (RCI)
 - % of patients who, by EOT, resemble patients who were not eligible in the first place (“Normative Threshold”; Saavedra et al., 2021, 2022a)

Funding

- NIMH grant R01MH124438 (Morgan-López, A.A.* & Lochman, J.E., MPIs)
- NIAAA grant R01AA025853 (Morgan-López, A.A.* & Hien, D.A., MPIs)
- IES grant R305A220244 (McDaniel, H.L., PI)
- NIDA grant R01DA023156 (Lochman, J.E., PI)
- NICHD grant R01HD079273 (Lochman, J.E., & Vernberg, E.M., MPIs)

Coping Power Collaborators

- Lissette M. Saavedra (RTI)
- Heather L. McDaniel (UVA)
- Stephen G. West (ASU, Free University of Berlin)
- Nicholas S. Ialongo (JHU)
- Catherine P. Bradshaw (UVA/JHU)
- Alexandra T. Tonigan (RTI/UNM)
- Nicole P. Powell (UA-T)
- Lixin Qu (UA-T/UNC-CH)
- Anna C. Yaros (RTI)
- John E. Lochman (UA-T)



Project Harmony Collaborators

- Lissette M. Saavedra (RTI)
- Denise A. Hien (Rutgers)
- Sudie E. Back (MUSC)
- Therese K. Killeen (MUSC)
- Sonya B. Norman (UCSD)
- Lesia M. Ruglass (CCNY)
- Skye S. Fitzpatrick (York/Toronto)
- Teresa López-Castro (CCNY)
- Chantel T. Ebrahimi (New School)



Advanced CSC/RCI Papers

Morgan-Lopez, A. A., Saavedra, L. M., Ramirez, D. D., Smith, L. M., & Yaros, A. C. (2022). Adapting the multilevel model for estimation of the reliable change index (RCI) with multiple timepoints and multiple sources of error. *International Journal of Methods in Psychiatric Research*, [1906]. <https://doi.org/10.1002/mpr.1906>

Morgan-López, A. A., Saavedra, L. M., McDaniel, H.L., West, S.G., Ialongo, N.S., Bradshaw, C.P., Tonigan, A.T., Montgomery, B.W., Powell, N.P., Qu., L., Yaros, A.C., & Lochman, J.E. (revise-and-resubmit) Beyond Jacobson and Truax: Estimation of Clinical Significance Trajectories in the Coping Power Intervention Using Measurement Error-Corrected Multilevel Modeling. Under review after invitation to resubmit to *Behavior Therapy*.

Saavedra, L. M., Morgan-López, A. A., Back, S. E., Patel, S. V., Hien, D. A., Killeen, T. K., ... Ruglass, L. M. (2022). Measurement error-corrected estimation of clinically significant change trajectories for interventions targeting comorbid PTSD and SUDs in OEF/OIF veterans. *Behavior Therapy*. 53(5), 1009-1023. <https://doi.org/10.1016/j.beth.2022.04.007>

Saavedra, L. M., Morgan-López, A. A., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Fitzpatrick, S., & Lopez-Castro, T. (2021). Putting the patient back in clinical significance: Moderated nonlinear factor analysis for estimating clinically significant change in treatment for posttraumatic stress disorder. *Journal of Traumatic Stress*, 34(2), 454–466. <https://doi.org/10.1002/jts.22624>



Thank you

Contact: Name | email: amorganlopez@rti.org