


ORIGINAL ARTICLE

Adapting the multilevel model for estimation of the reliable change index (RCI) with multiple timepoints and multiple sources of error

Antonio Alexander Morgan-Lopez¹  | Lissette Maria Saavedra¹ |
Derek D. Ramirez¹ | Luke M. Smith² | Anna Catherine Yaros¹

¹Behavioral Health Research Division, RTI International, Research Triangle Park, North Carolina, USA

²EL Futuro, Durham, North Carolina, USA

Correspondence

Antonio Alexander Morgan-Lopez, Behavioral Health Research Division, RTI International, Research Triangle Park, 3040 Cornwallis Rd, NC, 27709-2194, USA.

Email: amorganlopez@rti.org

Funding information

National Institute of Justice, Grant/Award Number: 2018-ZD-CX-0001

Abstract

Objective: One of the primary tools in the assessment of individual-level patient outcomes is Jacobson and Truax, (1991's) Reliable Change Index (RCI). Recent efforts to optimize the RCI have revolved around three issues: (a) extending the RCI beyond two timepoints, (b) estimating the RCI using scale scores from item response theory or factor analysis and (c) estimation of person- and time-specific standard errors of measurement.

Method: We present an adaptation of a two-stage procedure, a measurement error-corrected multilevel model, as a tool for RCI estimation (with accompanying Statistical Analysis System syntax). Using DASS-21 data from a community-based mental health center ($N = 379$), we illustrate the potential for the model as unifying framework for simultaneously addressing all three limitations in modeling individual-level RCI estimates.

Results: Compared to the optimal-fitting RCI model (moderated nonlinear factor analysis scoring with measurement error correction), an RCI model that uses DASS-21 total scores produced errors in RCI inferences in 50.8% of patients; this was largely driven by overestimation of the proportion of patients with statistically significant improvement.

Conclusion: Estimation of the RCI can now be enhanced by the use of latent variables, person- and time-specific measurement errors, and multiple timepoints.

KEYWORDS

modelling, statistics, psychometrics

1 | INTRODUCTION

Despite the focus on the treatment and monitoring of progress of individual patients in community settings, the overwhelming focus in clinical research contexts is often the examination of group-averaged

differences in changes over time across treatment arms. However, this focus on averaged changes over time in outcomes has been cited by clinicians, consumers, and third-party payers as one of many reasons why results from treatment outcome studies have had limited impact on intervention uptake (Ogles et al., 2001; Saavedra et al., 2019, 2021).

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. International Journal of Methods in Psychiatric Research published by John Wiley & Sons Ltd.

In contrast, approaches to assessing clinically significant change (clinically significant change [CSC]; Kazdin, 1977; Kendall et al., 1999) place a focus on improvement (or deterioration) in individual patients. Papers describing many of these approaches expressed parallel sets of criticisms of in assessing group-level change in outcome studies because group-based outcomes assessment cannot assess the proportion of individuals who were – and were not – able to return to a normal level of functioning (Jensen & Corrales, 2017; Westen et al., 2004).

While multiple approaches to the assessment of CSC have been proposed (e.g., movement below a “normative” threshold), one commonly applied approach to CSC is the RCI; Jacobson & Truax, 1991). The RCI assesses whether a patient’s change over time on the outcome measure cannot be attributable to error. The formula for the RCI is:

$$RCI_{CTT} = \frac{d_i}{SEM_d} \quad (1)$$

where d_i is an individual-level difference score (e.g., end-of-treatment – baseline) and SEM_d is the standard error of measurement. Patients are then classified by the direction of the change and whether each patient’s RCI was greater than $|1.96|$ (i.e., statistically significant at $p < 0.05$): (a) significant improvement (i.e., decrease), (b) nonsignificant improvement, (c) significant deterioration (i.e., increase) and (d) nonsignificant deterioration. However, as noted by Wise (2004), a more liberal criterion for individual-level significance of $p < 0.20$ may be more reasonable. This is because the functional “ n ” for the analysis is the number of timepoints for each person and the extremely small within-person n can make the RCI susceptible to Type II errors (Rozenal et al., 2017).

1.1 | Extensions of the RCI: Beyond pre-post and beyond sum scores

The original formulation of the RCI continues to be used, including recent applications in *International Journal of Methods in Psychiatric Research* (e.g., de Beurs et al., 2019; Schennach et al., 2016; Vaganian et al., 2020). However, multiple limitations of the original conceptualization of the RCI have been identified, relating to both the estimation of d_i and SEM_d . Three particular concerns have been noted: (a) the RCI numerator (d_i) has been limited to two timepoints, (b) an improper measurement model structure is often imposed on the scale scores that underlie d_i and (c) misspecification of the sources of error in estimating the RCI denominator (SEM_d). Misspecification of any or all of these elements can lead to errors in inference regarding whether a patient has improved or deteriorated (Jabrayilov et al., 2016; Saavedra et al., 2021).

Speer and Greenbaum (1995) were among the first to tap the potential of the latent growth curve model (LGCM) in estimating the RCI beyond two timepoints. In the conventional (unconditional) LGCM, both fixed effects and random effects are specified. The fixed

effects capture the sample average level of the outcome at time equals 0 (typically baseline) and the sample average slope per unit of time; the random effects for each individual capture the deviation from the average intercept and slope, respectively. The specification of LCGM-based RCI model requires an adaptation such that the model will have fixed effects of 0 (i.e., conditional means of 0 for the intercept and slope). This ensures that each individual’s slope is estimated in “raw” form instead of the deviation of patients’ trajectories from the sample average. This is critical, because tests for the significance of each individual’s slope would then be in their undeviated form, capturing whether each individual’s change over time differed from 0 – which Lovaglio and Parabiaghi (2014) note is an empirical Bayes estimate of d_i across multiple timepoints. However, this multiple timepoint RCI takes into account only one source of error (i.e., “Level-1” residual) and assumes perfect reliability in outcome scores.

Other work on the RCI (e.g., Brouwer et al., 2013; Jabrayilov et al., 2016; Saavedra et al., 2021) has addressed the need for greater precision in estimating the outcome scale scores that are the basis of the RCI numerator (d_i), and person- and time-specific measurement errors in the RCI denominator ($SEM_{d_{iti}}$) under various advanced forms of factor analysis (FA) or item response theory (IRT) (FA/IRT). The overwhelming majority of recent applications have used sum scores (e.g., de Beurs et al., 2019; Schennach et al., 2016). This has persisted despite many articles criticizing their use (e.g., Campbell, 1960; Dorans, 2007; Sijtsma, 2009). Scores estimated under FA or IRT conveys different amounts of clinical “weight” with regard to the relation of each item to the underlying construct (i.e., factor loadings in FA, discrimination parameters in IRT; Bollen, 1989; Embretson & Reise, 2000).

Conversely, the psychometric model underlying total scores (all FA loadings/ IRT discrimination parameters constrained to equality; Andrich, 1978) often fails to fit psychiatric symptom data (He et al., 2014; McNeish & Wolf, 2020). As a result, using total scores versus FA/IRT scores in estimating RCIs can lead to high proportions of discrepancies between RCI inference groupings (e.g., significant improvement, non-significant improvement, etc.), where upwards of 25% of patients can have a discrepant RCI judgement because of the differences in scale score estimation method (Brouwer et al., 2013; Saavedra et al., 2021). However, this work has been done primarily in the two-timepoint situation.

1.2 | A unifying framework for multiple timepoint RCI w/multiple sources of error

Very recent work in the quantitative methods literature by Diakow (2013) and Wang et al. (2019) on a two-stage procedure for estimating growth models with multiple sources of error (i.e., prediction error, measurement error) shows potential as a unifying framework for estimating the RCI. First, in Stage 1 they estimate scale scores and scale score standard errors from a measurement model under advanced FA or IRT (Jabrayilov et al., 2016; Saavedra et al., 2021).

In Stage 2, they read in the external measures of observation-specific measurement error (e.g., squared standard errors from FA or IRT), correcting individual treatment outcome trajectories for heteroskedastic measurement error in a manner similar to reading in external measures of sampling variance in meta-analysis (Sheu & Suzuki, 2001). Notably, these quantitative innovations were developed well outside of clinical contexts and make no particular reference to the RCI. Such a model addresses long-standing calls for mitigation of multiple threats to the validity in estimating RCIs, particularly in cases where a patient has been incorrectly judged as having significant improvement when they are either not improving or potentially getting worse (Hsu, 1989; Saavedra et al., 2021).

1.3 | The present study

We present an application of the proposed multiple-timepoint, multiple-error source RCI model using treatment outcomes monitoring data on negative affectivity from a community-based mental health treatment center (Saavedra et al., invited for resubmission) as a vehicle for demonstrating the method. The focal application uses scale scores and scale score standard errors estimated under the moderated nonlinear factor analysis (MNLFA) model (Bauer, 2017; Bauer & Hussong, 2009); the MNLFA model accounts for differential item functioning (DIF) in both factor loadings and item thresholds across multiple, simultaneous predictors. Often, these predictors are used in DIF analysis to separate out differences that are due to measurement artifacts from “true” differences in the construct of interest. We then compare RCI model

inferences from this focal model against three other competitor models, with a focus on a “conventional” model that uses sum scores with no correction for measurement error. These comparisons will help identify the relative contributions of two factors that have been shown to contribute to inaccuracy in RCI inferences: misspecification in the measurement model underlying the scale scores and/or inaccurate corrections for unreliability of the measure.

2 | METHOD

Participants. Participants consisted of 379 individuals who presented to a community based mental health treatment center that provides bilingual and culturally-informed behavioral health treatment for underserved primarily Latino individuals and families between 2008 and 2018. Participants were assessed at treatment intake and subsequent 3-month, 6-month and 12-month follow-ups; all were treated with evidence-based interventions, most of which were grounded in cognitive behavioral principles. As part of a comprehensive assessment, participants were informed that diagnostic instruments are available in both Spanish and English and they may select the language they are most comfortable with. Demographic information is presented in Table 1 with differences across language use shown, with participants requiring the Spanish version of the DASS-21 significantly more likely to be female, older and more likely to have a trauma diagnosis as their primary presenting problem. Data may be made available with joint permission of the first, second, and fourth authors.

TABLE 1 Descriptive statistic

	English (n = 115)	Spanish (n = 264)	p-value
Female	64.30%	73.80%	0.06
Age	22.26 (7.41)	37.33 (11.75)	<0.0001
Anxiety	7.83%	5.68%	0.006
Depression	46.09%	43.56%	
Trauma	24.35%	40.53%	
Substance use	4.35%	1.89%	
Other psychiatric	17.39%	8.34%	
DASS-21 scores			
Total score baseline	23.35 (14.62)	30.73 (15.86)	
Total score 3 Month	19.13 (13.33)	21.40 (15.03)	
Total score 6 month	19.13 (12.51)	21.17 (15.34)	
Total score 12 month	20.62 (13.72)	18.94 (14.57)	
MNLFA baseline	-0.11 (0.96)	0.41 (1.02)	
MNLFA 3 month	-0.41 (0.92)	-0.22 (1.00)	
MNLFA 6 month	-0.34 (0.79)	-0.22 (1.04)	
MNLFA 12 month	-0.35 (0.95)	-0.37 (0.95)	

Abbreviation: MNLFA, moderated nonlinear factor analysis.

2.1 | Measures

Depression, Anxiety and Stress Scale – 21-item version (DASS-21). The DASS-21 English (Antony et al., 1998; Henry & Crawford, 2005) and Spanish (Daza et al., 2002) versions consists of three 7-item self-report scales taken from the full version of the DASS (42 items). In many applications, including among Hispanic immigrants (Camacho et al., 2016), a unidimensional/single-factor structure has been supported; dimensionality of the DASS-21 in this sample is assessed prior to model fitting. Items are rated using a 4-point Likert-type scale (0 = did not apply to me at all” to 3 = “applied to me very much or most of the time”) describing negative emotional states. There are three subscales-one for depression, anxiety and stress. Sound internal consistency, as well as convergent and discriminant validity have been observed for the DASS-21 under classical test theory analyses (McDonald, 1999); these properties may differ for score estimation under MNLFA.

Predictors. Spanish/English language, gender, age and timepoint indicators were included as predictors of DIF; language, gender and age were also used as predictors in outcomes analysis independent of DIF in MNLFA analyses.

2.2 | Data analysis

Tests for Model Fit: Single-Factor Models. Prior to fitting of MNLFA models, unidimensional FA models were estimated to assess whether scale scores could be defensibly estimated without strict local independence between DASS-21 items (i.e., essential unidimensionality) using weighted least squares estimation in *Mplus* version 8 (Muthén & Muthén, 1998-2017). Within tests of fit of a general single factor, we also fit a more restrictive model where factor loadings/discrimination parameters were constrained to equality to assess whether a model that assumes equal factor loadings would fit the data (Andrich, 1978; He et al., 2014; McNeish & Wolf, 2020).

MNLFA. First, an initial base, single factor MNLFA model was fit where the 21 DASS symptoms were modeled assuming no DIF across language, time, gender or age. Next, a series of 21 models were fit for each item/symptom where thresholds and loadings were tested to see if they varied across language, time, gender or age (i.e., DIF). For each model, including the final MNLFA scoring model, DIF predictors were centered so that the interpretation of the “main” item parameters is at the mean levels across language, time, gender and age; DIF parameters for each predictor (if present) are interpreted as the deviation in the item parameter per 1 unit difference in the predictor. The final MNLFA scoring model generates latent negative affectivity scores that take into account any identified DIF across all symptoms. The final MNLFA scoring model generates (a) latent negative affectivity severity scores and (b) person- and time-specific standard errors of measurement which are output from *Mplus*; these were to be read into Statistical Analysis System (SAS) for outcomes analysis. Wang et al. (2019) and Zhang and Wang (2021) notes the advantages of separating the process of measurement modeling/scale score

estimation from outcomes analysis, especially with large numbers of repeated measures and/or many indicators of a latent construct.

Local Reliability. Local reliability (Embretson & Reise, 2000) was calculated by first extracting the test information function (TIF) values from the PLOT function in *Mplus* then converting the TIF to reliability values that are specific to different levels of underlying negative affectivity severity using the formula $(1-[1/TIF])$.

2.3 | Reliable change: Individual-level random effects and standard errors of measurement

To implement the RCI accounting for multiple timepoints and time- and person-specific measurement errors, an initial model under a conventional longitudinal (linear) multilevel model structure was estimated (see e.g., MacCallum et al., 1997; note that the final model structure was quadratic in nature):

$$Y_{it} = \beta_{0i} + \beta_{1i}(Time_{it}) + e_{it} \quad (2)$$

where, for the within-individual level, Y_{it} is the score on the outcome (sum score or FA/IRT score) measured for individual i at time t , β_{0i} is the random intercept for each individual at time = 0 (i.e., each person's baseline estimate), β_{1i} is the random slope-over-time for each individual and e_{it} is a residual term capturing Y_{it} 's deviation from each individual's predicted outcome trajectory. The between-individual model for the random effects β_{0i} and β_{1i} is presented here:

$$\beta_{0i} = g_{0i} \quad (3)$$

$$\beta_{1i} = g_{1i} \quad (4)$$

Typically, a growth model will have a fixed effect component that is (when no covariates are present), the average intercept and slope (β_{00} and β_{10}), respectively. However, in this model, fixed effects were set to 0 so that g_{0i} and g_{1i} are not interpreted as deviations from the mean intercept and slope but as their “raw” intercept and slope values. From this model, variance components (i.e., $Var[g_{0i}]$, $Var[g_{1i}]$, $Cov[g_{0i}, g_{1i}]$, $Var[e_{it}]$) are saved and used as fixed values in the next phase of analysis. g_{1i} from Equation (4), being the estimated individual slope over time, is the multiple timepoint equivalent of d_i in the RCI formula. As such, the standard error for each individual's estimate of g_{1i} could be used to test whether or not each person's slope was significantly different from 0. Speer and Greenbaum (1995) proposed exactly this as the RCI in multilevel models. However, to correct g_{1i} and its standard error for time- and person-specific measurement error, an additional term needed to be added to the model in the next phase of analysis (Diakow, 2013).

A second multilevel model was specified that has an additional error component:

$$Y_{it} = \beta_{0i} + \beta_{1i}(Time_{it}) + e_{it} + s_{it} \quad (5)$$

Equation (5) now includes a term, s_{it} , read into the analysis a fixed measurement error component, as the square of each

individual's standard error of measurement (SEM^2_{it}). The multilevel model is re-estimated, but with (a) the variance components ($Var[g_0]$, $Var[g_1]$, $Cov[g_0, g_1]$, $Var[e_{it}]$) constrained to their values from the first multilevel model and (b) values for s_{it} read in from the MNLFA scale score output, similar to reading in fixed study-level variance values in a meta-analysis (Diakow, 2013; Sheu & Suzuki, 2001); while s_{it} will vary for each individual and each timepoint under MNLFA, it is assumed constant for all individuals across all timepoints for total scores, using the formula for conversion of Cronbach's α to a (squared) standard error of measurement (McDonald, 1999).

With this additional term, estimates of g_1 and its standard error (and corresponding significance tests) now take into account multiple timepoints, residual error (e_{it}) and measurement error in the scale score (s_{it}). The resulting output (dataset) from the multilevel model will have each individual's estimate for g_1 and a corresponding RCI grouping can be generated based on each person's significance test for whether their level of change (improvement or deterioration over time) was statistically significant at $p < 0.20$ (Wise, 2004). Mplus MNLFA analysis syntax can be found in Bauer (2017). SAS Proc Mixed syntax for RCI analysis for multiple timepoints and multiple error sources is included in Supplementary Material for this paper but can be executed in any multilevel software that allows variance components to be constrained to pre-determined values (see e.g., Rabe-Hesketh et al., 2004).

3 | RESULTS

3.1 | Preliminary tests of model fit

The conventional single-factor model for DASS-21 items fit the data well, comparative fit index (CFI) = 0.97, root mean square error of approximation (RMSEA) = 0.054, 95% confidence (CI; 0.051, 0.057), meeting the standard for essential unidimensionality (Millsap & Kwok, 2004). A model with equality constraints on factor loadings fit comparatively worse against the conventional single-factor model ($\Delta\chi^2[20] = 679.75$, $p < 0.001$). Thus, a psychometric model that assumes that the DASS-21 items have equal weight (i.e., total scores) is likely to produce significant bias in scale scoring, individual trajectory estimation and subsequent misclassification of RCI inferences (Saavedra et al., 2021).

Moderated Non-Linear FA (MNLFA). A final model then included intercept and slope DIF parameters (above-and-beyond "true" differences in latent negative affectivity) that remained significant when included across all symptoms from interim DIF models. Item parameters from the final model are presented in Tables 2 and 3. The four items that showed no DIF across any of the factors examined, and served as empirical anchor items were: Items 8 ("...nervous energy"), 12 ("...found it difficult to relax"), 15 ("...felt I was close to panic") and 16 ("...unable to become enthusiastic..."). The final DASS-21 scoring model, with DIF incorporated across all other items under a partial invariance model (Millsap & Kwok, 2004), fit significantly better than the base model $\chi^2(46) = 838.69$, $p < 0.0001$. This model

was the model under which MNLFA scale scores and standard errors were estimated, output from Mplus, and used in primary multiple-timepoint, multiple error source RCI models.

Local Reliability. Figure 1 shows the local reliability curves (Chiesi et al., 2017; Morgan-Lopez et al., 2020) for the DASS-21 scores separately for both languages. Reliability values for the DASS-21 scores, as expected, remain above 0.90 throughout a considerable range of the latent negative affectivity scores (between 1.5 SDs below and 2.5 SDs above the mean). By comparison, conventional reliability under Cronbach's α was 0.955 for Spanish and 0.947 for English.

3.2 | Multilevel RCI models: Output and classification of individual trajectories

Preliminary models suggested significant deceleration in treatment trajectories for both DASS-21 MNLFA scale scores ($b = 0.10$ [0.02], $t = 6.71$, $p < 0.0001$) and total scores ($b = 1.56$ [0.25], $t = 6.35$, $p < 0.0001$), thus quadratic effects were included. Final multilevel RCI model random effect covariance parameters for each model (across scoring method and measurement error correction method) are shown in Table 4, with accompanying SAS Proc MIXED syntax contained in Supplementary Material 1. From the individual-level linear slope values, output as a SAS dataset, patients were classified across all 4 models based on both (a) whether their individual linear slope value was significant/non-significant at $p < 0.20$ and (b) whether it was positive (deterioration) or negative (improvement). The RCI classification percentages for patients based on the MNLFA scale scores with externally-entered (squared) standard errors of measurement were: 21.5% of patients had significant improvement, 51.6% of patients with non-significant improvement, 24.5% of patients with non-significant deterioration and 2.3% of patients with significant deterioration. Subsequent multinomial logistic regression analysis of the RCI groupings showed significant differences in classification only for language grouping among all covariates ($\chi^2[3] = 8.85$, $p = 0.03$), with the bulk of the differences driven by higher proportions of Spanish speakers with significant improvement (25.8%) compared to English speakers (11.7%).

3.3 | Tests for agreement between scoring methods

Tests for classification agreement in dependent contingency tables (Bowker, 1948) were conducted on the classifications for the RCI across each combination of scale scoring method (MNLFA, total scores) and handling of measurement error (modeled, not modeled) to assess the level of (dis)agreement between RCI inferences across methods. Given the optimal fit of the (MNL) FA measurement model, the RCI based on the MNLFA scores with measurement error correction is deemed the "Reference" model.

"Reference" RCI versus RCI under MNLFA without Measurement Errors. The test for agreement between the Reference RCI against MNLFA without measurement error correction assesses the

TABLE 2 MNLFA general item parameters

Item	Factor loading	Threshold (0-1)	Threshold (1-2)	Threshold (2-3)
Hard to wind down	1.63	-2.64	-0.26	1.83
Dryness of the mouth	1.17	-0.74	0.63	2.15
Couldn't experience positive feeling	2.07	-1.46	0.85	3.14
Difficulty breathing	1.53	-0.14	1.44	3.5
Lack of initiative	1.62	-1.85	0.11	2.07
Overreacting	1.58	-1.66	0.14	2.18
Trembling hands	1.74	-0.52	1.03	2.86
Nervous energy	2.24	-2.21	0.14	2.32
Worried about panic/fool	1.91	-0.64	0.88	2.68
Nothing to look forward to	2.29	0.63	2.43	4.26
Agitated	1.93	-1.47	0.77	3
Difficult to relax	2.23	-2.44	-0.09	2.4
Blue	2.45	-2.66	-0.11	2.36
Intolerant	1.74	-1.26	0.75	2.79
Close to panic	2.54	-0.22	1.8	4.21
Unable to become enthusiastic	2.41	-0.92	1.36	3.78
Worthless	2.18	-0.33	1.25	3.17
Touchy	1.74	-1.35	0.54	2.58
Aware of increased heart rate	1.82	-0.66	0.99	3.03
Scared	2.17	-0.68	1.03	3.1
Life was meaningless	2.18	0.69	2.24	4.08

Abbreviation: MNLFA, moderated nonlinear factor analysis.

extent to which the RCI inferences deviate when the scores from the correct measurement model are used but perfect reliability is assumed. The test for agreement was significant, $S(6) = 15.75$, $p = 0.015$; however, the Weighted Cohen's Kappa of 0.84 (CI: 0.79, 0.89) suggesting strong agreement between the methods, with 88% marginal agreement between RCI classifications. The marginal percentage of patients who achieved statistically significant improvement (SSI) using the MNLFA scale scores without measurement error correction (25.9%) was slightly higher than the Reference RCI (21.5%).

"Reference" RCI versus RCI using Total Scores and Classical Test Theory Measurement Error Correction. The test for agreement between the Reference RCI against the RCI using total scores with measurement error correction based on Cronbach's α (McDonald, 1999) assesses the extent to which the RCI inferences deviate when (a) the scores from a worse fitting measurement model are used and (b) measurement error correction is modeled but is also incorrectly assumed to be the same for all patients at all timepoints. The test for agreement was significant, $S(6) = 42.28$, $p < 0.001$, with weak agreement based on the Weighted Cohen's Kappa of 0.55 (CI: 0.48, 0.61), with 62.6% agreement between RCI classifications, suggesting that greater than one in 3 patients would have been misclassified using this model. The significant test for agreement was driven by

differences in the marginal percentages of patients who achieved SSI using the total scores with measurement error correction (35.2%) versus the Reference RCI (21.5%).

"Reference" RCI versus RCI using Total Scores without Measurement Error Correction. The test for agreement between the Reference RCI against the RCI using total scores without measurement error correction assesses the extent to which the RCI inferences would deviate from the model where a multilevel model is used to estimate the RCI but total scores are used in a conventional manner (Lovaglio & Parabiaghi, 2014; Speer & Greenbaum, 1995). The test for agreement was significant, $S(6) = 153.38$, $p < 0.001$, with a Weighted Cohen's Kappa value that would be considered less than the convention for weak agreement, 0.36 (CI: 0.30, 0.43), with 49.2% agreement between RCI classifications, suggesting that half of patients would have been misclassified using a "conventional" measurement model structure total scores treated as perfectly reliable. The significant test for agreement was driven largely by differences in the marginal percentages of patients who achieved SSI using the total scores without measurement error correction (49.8%) versus the Reference RCI (21.5%; see Table 5). 90 patients (26.3% of the sample) were judged as having SSI using total scores that were judged as having non-significant improvement using MNLFA scores

TABLE 3 MNLFA differential item functioning (DIF) parameters

Item	Lang. τ DIF	Gender τ DIF	Age τ DIF	3 Month τ DIF	6 Month τ DIF	12 Month τ DIF	Lang. λ DIF	Gender λ DIF	Age λ DIF	3 Month λ DIF	6 Month λ DIF	12 Month λ DIF
Hard to wind down	0.63											
Dryness of the mouth	0.23	-0.34	0.13								0.45	
No positive feeling												
Difficulty breathing												
Lack of initiative	0.51											
Overreacting							0.392		-0.14			
Trembling hands	0.63			0.26								
Nervous energy												
Worried about panic/fool	-0.61		-0.15			0.38						
Nothing to look forward to	-1.74	-0.39										
Agitated	-1.35			0.25						0.38		
Difficult to relax												
Blue	0.47	0.28	0.2	-0.39	-0.37	-0.5						
Intolerant	0.5		-0.22				0.334					
Close to panic												
Unable to become enthusiastic												
Worthless												
Touchy								0.31				
Aware of increased heart rate	0.415	-0.27				0.42						
Scared	0.289											
Life was meaningless	-0.87				-0.32							

Note: τ = item threshold. λ = factor loading.

Abbreviations: DIF, differential item functioning; MNLFA, moderated nonlinear factor analysis.

with measurement error corrections. Another 67 patients (19.7% of the sample) who were judged as having non-significant improvement using total scores had non-significant *deterioration* based on MNLFA scores with measurement error corrections.

4 | DISCUSSION

The overemphasis on examining averaged changes over time in assessing treatment efficacy in psychiatric research settings stands in contrast to the focus on progress or deterioration among individual patients in community practice. Often, these emphases can lead to somewhat contradictory conclusions, in cases where the change over time in treatment outcomes suggests significant/meaningful improvements *on average*, yet a non-trivial proportion of individual patients fail to improve or are getting worse (Jensen &

Corralejo, 2017; Saavedra et al., 2021; Westen et al., 2004). Interest in the assessment of individual-level clinical change has been longstanding (see e.g., Kazdin, 1977) but has generally coalesced around the RCI developed by Jacobson and Truax (1991) which continues to be used in its original form.

We present a form of RCI estimation that addresses several noted limitations of the original RCI under an adapted measurement error-corrected multilevel model. Estimation of RCIs under multilevel models with different assumptions about DASS-21 scale scores and reliabilities suggested some serious practical consequences for what can happen when the measurement model that underlies total scores does not fit the data. For example, when the focal RCI model (MNLFA scores, person- and time-specific measurement errors) was compared against what was essentially a "standard practice" model (total scores, no correction for measurement error), the classification accuracy of individual trajectories into RCI groupings (e.g., SSI, etc.) was

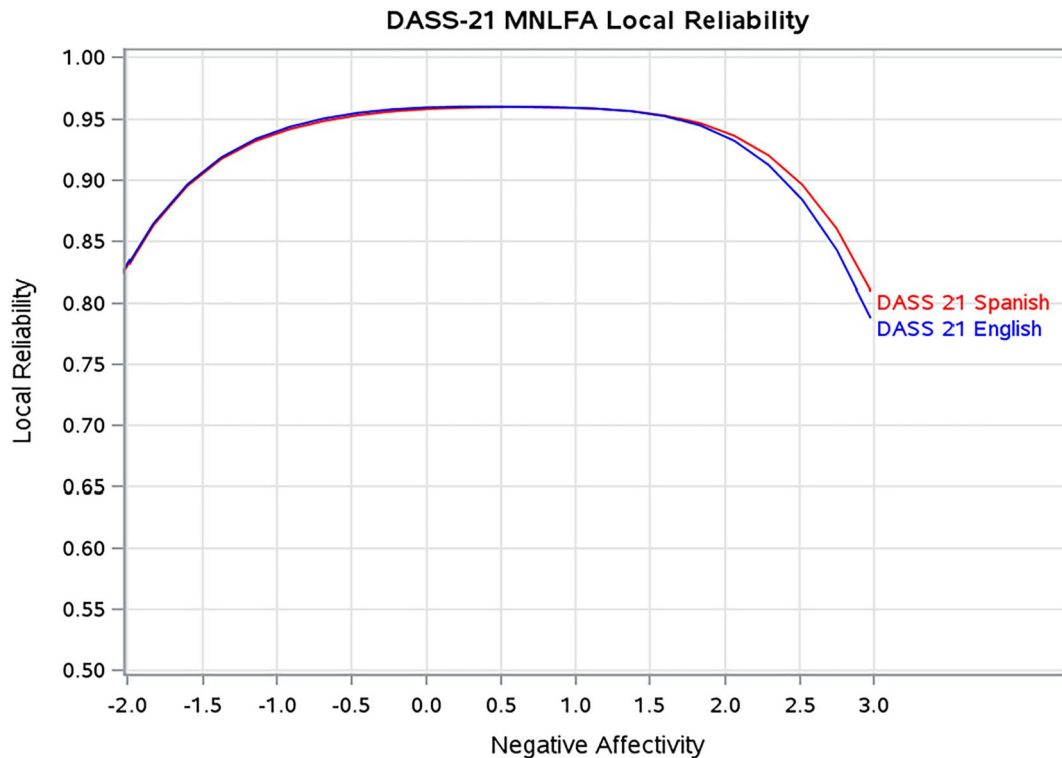


FIGURE 1 Local reliability plots

TABLE 4 Reliable change index (RCI) multilevel modeling variance components

Multilevel modeling variance components	MNLFA scale scores (no measurement error)	MNLFA Scale scores with person- and time-specific SE ² s	Sum scores (no measurement error)	Sum scores with single SEM ²
V (intercept)	0.74 (0.10)	0.74	1019.28 (80.80)	1019.28
Cov (int, time)	-0.23 (0.09)	-0.23	-284.20 (38.97)	-284.20
V (time)	0.52 (0.15)	0.52	148.98 (31.98)	148.98
Cov (int, time ²)	0.02 (0.02)	0.02	48.12 (9.53)	48.12
Cov (time, time ²)	-0.11 (0.04)	-0.11	-31.68 (8.11)	-31.68
V (time ²)	0.02 (0.02)	0.02	7.40 (2.13)	7.40
Squared standard error(s) of measurement	N/A	0.05-0.35*	N/A	11.90
Within-person residual	0.38 (0.05)	0.38	77.36 (8.79)	77.36

Note: Estimated variance components from the models with no measurement error were used as fixed values in the models where (squared) standard errors of measurement were incorporated. *Range of squared standard errors of measurement (rounded to the nearest 0.05).

Abbreviations: MNLFA, moderated nonlinear factor analysis; RCI, Reliable Change Index.

(a) no better than a coin flip and (b) suggested that more than twice the number of patients (49.8% vs. 21.5%) would have been classified as having SSI using a scale scoring model that had already failed to fit the data in the measurement model steps. This could have significant practical implications such as, for example, discharging a patient who appeared to be improving that may not have been improving at all (Sinharay & Haberman, 2014). Other RCI analyses that were conducted to disentangle how much of the discrepancy in the RCI was attributable to the scale scoring and the handling of measurement error suggested that the method of scale scoring had much more

impact on RCI inferences than did handling of measurement error. This is not surprising, given the metrics on both conventional reliability (>0.94) and local reliability (above 0.85 through most of the range of latent negative affectivity) were strong. Nevertheless it cannot be always assumed that measurement error would have minimal impact and, in fact, 12% of patients would have had a different RCI inference using the MNLFA scores but treating them as perfectly reliable.

These results overall also point us to a clinically relevant case-in-point concerning the tension between group- and individual-level

TABLE 5 RCI agreement table

RCI from DASS-21 MNLFA scores (with measurement error correction)	RCI from DASS-21 sum scores (no measurement error correction)				Total (marginal percentages)
	Significant deterioration	Non-significant deterioration	Non-significant improvement	Significant improvement	
Significant deterioration	3	4	0	1	82.35
Non-significant deterioration	0	11	67	5	8324.34
Non-significant improvement	0	5	81	90	17651.61
Significant improvement	0	0	0	74	7421.70
Total (marginal percentages)	30.88	205.87	14843.40	17049.85	341100.00

Abbreviations: DASS, Depression anxiety and stress scale; MNLFA, moderated nonlinear factor analysis; RCI, Reliable Change Index.

inferences. A fairly sizeable effect size was observed through 12-month follow-up, with similar overall effect sizes (pooled $d > |0.5|$) for both total score and MNLFA scoring methods; compatible group-averaged statistics are common across scoring methods given the property of stochastic ordering of total scores and FA/IRT scores (Ellis & Junker, 1997; van der Ark, 2005). However, individual-level effects as estimated under the RCI showed that ~26% of patients had treatment outcome deterioration under MNLFA scoring but only ~5% using total scores. This discrepancy points out two issues. First, the discrepancy in findings for the RCI across scoring methods points to the notion that differences in scoring methods are more likely to impact individual scores and trajectories rather than overall sample statistics and group-averaged findings (Hanson et al., 2001; Kim & DeCarlo, 2016). Second, having a non-trivial proportion of patients who are shown to get worse via the RCI illustrates the potential perils of generalizing group-level averages to individual-level treatment outcomes (Jacobson & Truax, 1991; Kazdin, 1977; Kendall et al., 1999). Effect sizes provide little-to-no information regarding whether any individual patient has actually moved from a clinical to non-clinical state and/or had reductions in outcomes at a level greater than would be expected by chance (Jensen & Corrales, 2017). Thus, using both group- and individual-level outcome metrics may serve the dual purpose of showing overall efficacy while also giving patients and providers individual-level outcome information that is more relevant and specific to their interests in patient progress or deterioration (Saavedra et al., 2019, 2021).

4.1 | Limitations

There are two limitations in particular to note. First, these data were taken from patients seeking mental health services in a community-based mental health treatment center that were not part of a randomized controlled trial. While the outcome results point to clear overall average improvement over time in negative affectivity, at a rate much greater than would be expected by chance, no specific claim is made regarding the impact of *any specific treatment* or *mechanisms of treatment action* on this improvement. The primary interests in this study were centered around joint measurement and individual trajectory modeling (Wang et al., 2019); all of the well-

known limitations regarding treatment effects in single-group designs apply to this study (Campbell & Stanley, 1963; Shadish et al., 2002).

Second, the lack of a no-treatment comparison condition does present some concerns regarding the decomposition of DIF effects for assessment wave. While not discussed in the Results, there were a number of effects showing significant DIF by assessment wave (15 of 84%, 18% of DIF tests were significant) using three dummy indicators where baseline was the reference timepoint. Although the MNLFA model properly controlled for this DIF in scale score estimation (and standard sum scoring does not), what remains unclear is whether this assessment wave DIF was due to treatment, generalized assessment reactivity (Donovan et al., 2012) or both. While these effects cannot be disentangled in this study, other studies that are actual RCTs can better disentangle these effects when using the multiple timepoint, multiple error source RCI model (Saavedra et al., 2021, under review).

Third, RCI inferences were based on the linear slope in the presence of estimated quadratic effects for each person. Of the ~21% of the sample judged as having SSI, 75% of those patients had statistically significant *deceleration*, suggesting that their estimated improvements slowed beyond 3 months. While this may not have implications for the proposed RCI methodology per se, it does beg the question of how to assess whether the person's trajectory of change is significantly different from 0 in the presence of non-linearity. A potential option may be to shift the perspective from whether the slope estimate is different from 0 to whether the model-implied score at a particular timepoint from the multilevel model is significantly different from baseline. This is done, for example, in incorporating non-linearities in estimating longitudinal effect sizes (Feingold, 2019). However, such extensions are beyond the scope of this paper.

5 | CONCLUSION

Characterizing treatment progress requires precision in measurement and a balance between the overall treatment progress of groups against the individual patient trajectories. The present study illustrates the utility of an adaptation of Wang et al. (2019)'s two-stage framework for estimating the RCI (Jacobson & Truax, 1991). This approach unifies all of the very recent work on the RCI that has

either focused on (a) extension of the RCI beyond pre-post situations, mitigating measurement bias in the RCI by proper modeling of scale scores or (c) recognizing heteroskedasticity in measurement error across patients and/or time. This study also illustrates the long-held notion that group-level findings are not necessarily accurate reflections of individual-level treatment outcomes and prognosis. From a practical application standpoint, we would encourage the development of mobile apps to score underlying psychiatric severity where the MNLFA item parameters (e.g., Tables 2 and 3) would be “under the hood” of the app. This is fast becoming the case with computerized adaptive psychiatric scoring modules such as Patient-reported outcomes measurement information system (Cella et al., 2010) and CAT-MH (Gibbons & deGruy, 2019) that are grounded in IRT. Such apps would allow clinicians to take advantage of the benefits of these advanced scoring methodologies, even if they are not well-versed in advanced psychometrics. This would contribute greatly to improving accuracy and precision of routine clinical assessment without burdening the end user with complex scale scoring methodology. Psychiatric researchers are encouraged to incorporate modern methodologies of assessing item responses and scale score estimation into clinical applications as they strive for greater accuracy in assessment both for assessing group-level and individual-level treatment outcomes.

ACKNOWLEDGEMENTS

The work presented in this manuscript was supported by grants from the National Institute of Justice (NIJ grant 2018-ZD-CX-0001, Saavedra, L.M., PI) and funding from the RTI Fellows Program (Morgan-Lopez, A.A., PI).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author with permission from the fourth author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Antonio Alexander Morgan-Lopez  <https://orcid.org/0000-0003-4706-9964>

REFERENCES

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573. <https://doi.org/10.1007/BF02293814>
- Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, 10, 176–181.
- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526.
- Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, 14(2), 101–125. <https://doi.org/10.1037/a0015583>
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43(244), 572–574.
- Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research*, 23(5), 489–501.
- Camacho, Á., Cordero, E. D., & Perkins, T. (2016). Psychometric properties of the DASS-21 among Latina/o college students by the US-Mexico border. *Journal of Immigrant and Minority Health*, 18(5), 1017–1023.
- Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15, 546–553.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., ... Hays, R., PROMIS Cooperative Group. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63, 1179–1194.
- Chiesi, F., Primi, C., Pigliatelli, M., Ercolani, S., della Staffa, M. C., Longo, A., Mecocci, P., & Mecocci, P. (2017). The local reliability of the 15-item version of the Geriatric Depression Scale: An item response theory (IRT) study. *Journal of Psychosomatic Research*, 96, 84–88.
- Daza, P., Novy, D. M., Stanley, M. A., & Averill, P. (2002). The depression anxiety stress scale-21: Spanish translation and validation with a hispanic sample. *Journal of Psychopathology and Behavioral Assessment*, 2, 195–205.
- de Beurs, E., Carlier, I. V., & van Hemert, A. M. (2019). Approaches to denote treatment outcome: Clinical significance and clinical global impression compared. *International Journal of Methods in Psychiatric Research*, 28, e1797.
- Diakow, R. P. (2013). *Improving explanatory inferences from assessments*. Doctoral Dissertation, University of California. <https://escholarship.org/content/qt4fc64449/qt4fc64449.pdf>
- Donovan, D. M., Bogenschutz, M. P., Perl, H., Forcehimes, A., Adinoff, B., Mandler, R., Oden, N., & Walker, R. (2012). Study design to examine the potential role of assessment reactivity in the screening, motivational assessment, referral, and treatment in emergency departments (SMART-ED) protocol. *Addiction Science & Clinical Practice*, 7(1), 16.
- Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(1), 85–94.
- Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, 62, 495–523.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.
- Feingold, A. (2019). Time-varying effect sizes for quadratic growth models in multilevel and latent growth modeling. *Structural Equation Modeling*, 26(3), 418–429.
- Gibbons, R. D., & deGruy, F. V. (2019). Without wasting a word: Extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CAT-MH). *Current Psychiatry Reports*, 21, 1–9.
- Hanson, B. A., Harris, D. J., Pommerich, M., Sconing, J. A., & Yi, Q. (2001). Suggestions for the Evaluation and Use of Concordance Results. ACT Research Report Series.
- He, Q., Glas, C. A., & Veldkamp, B. P. (2014). Assessing impact of differential symptom functioning on post-traumatic stress disorder (PTSD) diagnosis. *International Journal of Methods in Psychiatric Research*, 23, 131–141.
- Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and

- normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 44, 227–239.
- Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment*, 11, 459–467.
- Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40, 559–572.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19.
- Jensen, S. A., & Corrales, S. M. (2017). Measurement issues: Large effect sizes do not mean most people get better—clinical significance and the importance of individual results. *Child and Adolescent Mental Health*, 22(3), 163–166.
- Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1(4), 427–452.
- Kendall, P. C., Mars-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 285–299.
- Kim, Y., & DeCarlo, L. T. (2016). *Evaluating equity at the local level using bootstrap tests*. The College Board. Research Report 2016–4, Rep. No. 4.
- Lovaglio, P. G., & Parabiaghi, A. (2014). Assessment of meaningful change in routine outcome measurement (ROM) with a combination of a longitudinal and a 'classify and count' approach. *Quality and Quantity*, 48, 2479–2499.
- MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32, 215–253.
- McDonald, R. P. (1999). *Test theory: A unified approach*. Psychology Press.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287–2305.
- Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93–115.
- Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Fitzpatrick, S., López-Castro, T., & Patock-Peckham, J. A. (2020). Estimation of equable scale scores and treatment outcomes from patient-and clinician-reported PTSD measures using item response theory calibration. *Psychological Assessment*, 32, 321–335.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review*, 21, 421–446.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual*. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160. <https://biostats.bepress.com/ucbbiostat/paper160>
- Rozental, A., Magnusson, K., Boettcher, J., Andersson, G., & Carlbring, P. (2017). For better or worse: An individual patient data meta-analysis of deterioration among participants receiving Internet-based cognitive behavior therapy. *Journal of Consulting and Clinical Psychology*, 85, 160–177.
- Saavedra, L. M., Morgan-López, A. A., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Lopez-Castro, T., & 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, 454–466.
- Saavedra, L. M., Morgan-López, A. A., Smith, L. M., Hayes, M. B., Dean, K., Siu, K., Ramirez, D. D., & Yaros, A. C. Graded response item response theory in the development of a shortened DASS-21 for use in telehealth settings in patients with primary affective disorders. Invited for resubmission to. *Journal of Clinical Psychology*, (under review).
- Saavedra, L. M., Morgan-Lopez, A. A., Yaros, A. C., Buben, A., & Trudeau, J. V. (2019). *Provider resistance to evidence-based practice in schools: Why it happens and how to plan for it in evaluations* (p. 8638). RTI Press.
- Schennach, R., Möller, H. J., Obermeier, M., Seemüller, F., Jäger, M., Schmauss, M., Laux, G., Pfeiffer, H., Naber, D., Schmidt, L. G., Gaebel, W., Klosterkötter, J., Heuser, I., Maier, W., Lemke, M. R., Rütger, E., Klingberg, S., Gastpar, M., Musil, R., ... Riedel, M. (2016). Challenging the understanding of significant improvement and outcome in schizophrenia—the concept of reliable and clinically significant change methods. *International Journal of Methods in Psychiatric Research*, 25(1), 3–11.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Sheu, C. F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers*, 33, 102–107.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120.
- Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, 33(1), 23–35.
- Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, 63, 1044–1048.
- Vaganian, L., Bussmann, S., Gerlach, A. L., Kusch, M., Labouvie, H., & Cwik, J. C. (2020). Critical consideration of assessment methods for clinically significant changes of mental distress after psycho-oncological interventions. *International Journal of Methods in Psychiatric Research*, 29, e1821.
- vander Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, 70, 283–304.
- Wang, C., Xu, G., & Zhang, X. (2019). Correction for item response theory latent trait measurement error in linear mixed effects models. *Psychometrika*, 84, 673–700.
- Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, 130, 631–663.
- Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment*, 82, 50–59.
- Zhang, X., & Wang, C. (2021). Measurement bias and error correction in a two-stage estimation for multilevel IRT models. *British Journal of Mathematical and Statistical Psychology*. <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12233>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: 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*, 31(2), e1906. <https://doi.org/10.1002/mpr.1906>