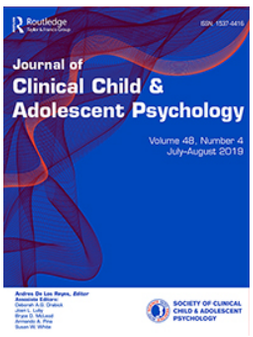

Routledge
Taylor & Francis Group

Journal of Clinical Child & Adolescent Psychology

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

Data Harmonization: Establishing Measurement Invariance across Different Assessments of the Same Construct across Adolescence

Fanita A. Tyrell, Tuppett M. Yates, Keith F. Widaman, Chandra A. Reynolds & William V. Fabricius

To cite this article: Fanita A. Tyrell, Tuppett M. Yates, Keith F. Widaman, Chandra A. Reynolds & William V. Fabricius (2019) Data Harmonization: Establishing Measurement Invariance across Different Assessments of the Same Construct across Adolescence, Journal of Clinical Child & Adolescent Psychology, 48:4, 555-567, DOI: 10.1080/15374416.2019.1622124

To link to this article: https://doi.org/10.1080/15374416.2019.1622124

Journal of Clinical Child & Adolescent Psychology, 48(4), 555–567, 2019

ISSN: 1537-4416 print/1537-4424 online
DOI: https://doi.org/10.1080/15374416.2019.1622124

Routledge
Taylor & Francis Group

Check for updates

EVIDENCE BASE UPDATE

Data Harmonization: Establishing Measurement Invariance across Different Assessments of the Same Construct across Adolescence

Fanita A. Tyrell
Institute of Child Development, University of Minnesota Twin Cities

Tuppett M. Yates
Department of Psychology, University of California Riverside

Keith F. Widaman
Graduate School of Education, University of California Riverside

Chandra A. Reynolds
Department of Psychology, University of California Riverside

William V. Fabricius
Department of Psychology, Arizona State University

Longitudinal measurement invariance is a major concern for developmental scholars who seek to evaluate the same underlying construct across time. Unfortunately, discontinuities in the expression of various psychological constructs, as well as essential changes in measurement that are necessitated by shifting developmental capacities and practice effects over time, make the task of establishing longitudinal invariance extremely difficult. Drawing on 5 waves of longitudinal data from 392 families (52% female; M_{age_W1} = 12.89, *SD* = .48; M_{age_W5} = 21.95, *SD* = .77; 199 European American and 193 Mexican American families), the current investigation sought to establish measurement invariance across developmentally appropriate changes in measures of depressive symptomatology from early adolescence through early adulthood. Using a combination of item parceling and the common and unique items from 2 assessment instruments for depressive symptoms, the data supported strong factorial invariance in youth's depressive symptoms across 5 waves of measurement. Findings suggest that traditional invariance approaches can be adapted to determine whether the same construct underlies different measurement instruments across time. This analytic strategy can allow researchers and clinicians to use more sophisticated techniques to understand changes in symptomatology regardless of changes in measurement or developmental capacity. Applying this approach to model patterns of depressive symptomatology from early adolescence to early adulthood has important clinical implications for elucidating periods when youth experience elevations in depressive symptoms and heightened needs for intervention services.

Address correspondence to Fanita A. Tyrell, Institute of Child Development, University of Minnesota, Twin Cities, 51 E River Rd, Minneapolis, MN 55455. E-mail: tyrel002@umn.edu

Measurement invariance entails the evaluation of whether the measurement parameters of latent variable indicators remain invariant across groups or occasions with respect to their factor loadings (i.e., weak invariance), intercepts (i.e., strong

invariance), and unique variances (i.e., strict invariance; Widaman, Ferrer, & Conger, 2010; Widaman & Reise, 1997) . *Longitudinal* measurement invariance gives researchers confidence that inferences about changes in participants' scores reflect changes in the construct of interest rather than changes in measurement and/or participant characteristics. The need to establish longitudinal measurement invariance poses unique challenges for developmental researchers because developmentally appropriate assessment protocols often necessitate changes in measurement as assessment strategies improve and/or as participants' capacities develop. Indeed, measurement change is required to model heterotypic continuity wherein manifest expressions of a construct are theorized to change, but its underlying developmental significance and function remain stable (Patterson, 1993; Sroufe & Jacobvitz, 1989). For example, whereas depressed mood may be expressed as irritability and guilt in young children, sadness and hopelessness characterize depression among adolescents and adults (Weiss & Garber, 2003). Unfortunately, traditional approaches to assess longitudinal measurement invariance (e.g., modeling associations of the same construct across multiple developmental periods) fall short when developmental research designs necessitate measurement change. Thus, the purpose of the current investigation was to demonstrate how traditional invariance approaches can be adapted to model both common and unique measurement items across development despite the use of different assessments to measure the same construct across time.

Traditional Measurement Invariance Approaches

Researchers who are able to assess a construct of interest using the same measurement instrument across time can readily evaluate longitudinal measurement invariance using initial factorial invariance procedures that were formulated by Meredith (1993) and later expanded and formalized by Widaman and Reise (1997). In this approach, hierarchically nested models characterized by increasingly strict constraints at each level are evaluated across configural, weak, strong, and strict forms of factorial invariance (Meredith, 1993; Widaman et al., 2010; Widaman & Reise, 1997). At the broadest level, configural factorial invariance evaluates whether the same pattern of fixed and free factor loadings is consistent across measurement occasions. If configural factorial invariance can be established, the next step is to test for weak factorial invariance, which entails determining whether the factor loadings for the latent construct of interest are invariant across time. For this reason, weak factorial invariance is also known as metric factorial invariance. Weak factorial invariance is necessary to demonstrate that the same underlying construct is being assessed across time. The third phase of traditional invariance analysis emphasizes strong factorial invariance, in which both the factor loadings and the intercepts of the observed variables are equivalent over time. Strong factorial invariance is necessary to support the examination of change in level, or growth, of a particular construct across time. Finally, strict factorial invariance is obtained by imposing additional equality constraints on the unique variances of the indicators across time, as well as ensuring equivalence in the factor loadings (i.e., weak factorial invariance) and intercepts (i.e., strong factorial invariance). Although it is possible to achieve strict factorial invariance in rare cases when constructs remain relatively stable in level and expression across time (e.g., intelligence, extroversion), developmental assessments rarely achieve this most stringent level of invariance.

When measures must change across time to capture a phenomenon of interest accurately, one of the foremost analytic approaches to evaluate invariance is to model the associations of the construct across time despite expected differences in measurement (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). However, it is problematic to formulate strong conclusions about the developmental continuity of a construct based strictly on observed associations between different measures because the magnitude or direction of these associations could be attributed to sources other than the construct itself, such as the use of shared informants or methods. More important, correlative approaches cannot evaluate change in the construct at the individual level, nor can they respond to recent calls for developmental investigations that differentiate interindividual (i.e., trait-invariant or between-person differences) from intraindividual (i.e., time-varying or within-person differences) patterns of change across time (Berry & Willoughby, 2016; Hamaker, Kuiper, & Grasman, 2015). Other strategies that employ standardization and regression approaches have similar limitations for evaluating intraindividual change (cf. McArdle et al., 2009).

During the last 10 years, item response theory (IRT) has emerged as a promising strategy to assess the same construct across time despite developmentally appropriate changes in measurement (Curran et al., 2008; Khoo, West, Wu, & Kwok, 2006; McArdle et al., 2009). A fundamental step toward establishing measurement invariance with IRT is vertical equating (Khoo et al., 2006; McArdle et al., 2009), in which common items are linked and equated on a common scale. Measurement invariance using IRT involves the consideration of two parameters for each item: the difficulty parameter (i.e., the difficulty of endorsing or passing the item) and the discrimination or slope parameter (i.e., the strength of the association between the item and the underlying construct; Steinberg & Thissen, 2013). To ensure measurement invariance, both the difficulty and discrimination parameters of linking items (i.e., items that are shared across measures) must be invariant across time, which means that items have to be equally difficult and discriminating across time, though the latent variable mean and variance may change across time (Khoo et al., 2006; McArdle et al., 2009). Although IRT is often used in education research to develop

computerized adaptive tests or measure student learning outcomes (e.g., Cook & Eignor, 1991; Khoo et al., 2006), typical IRT programs do not have the capability to examine dynamic and multidimensional psychological constructs over time with complex longitudinal models (Khoo et al., 2006). In recent years, scholars have shown that IRT models can be used to fit measurement models for select constructs tapping ability (i.e., vocabulary and memory; McArdle et al., 2009) or internalizing symptoms (Curran et al., 2008). However, the labor-intensive IRT process itself, as well as its limited utility for modeling complex age-related phenomena, have hindered its uptake among developmental researchers in search of an accessible approach for modeling developmental change over time in cases where assessments must also change.

Longitudinal Measurement Invariance with Changing Measures

Whether to mitigate practice effects, capitalize on emergent developmental capacities, or capture heterotypic patterns of continuity, change in measurement instruments is an essential feature of developmental research designs that follow participants across extended periods. Because developmental phenomena can change in form and/or function over time (Patterson, 1993; Sroufe, Egeland, & Carlson, 1999; Sroufe & Jacobvitz, 1989), the ideal approach to evaluating longitudinal measurement invariance must capture both the stability in the construct of interest and the qualitative shifts in measurements that may characterize the construct's expression over time. Although latent variable models can capture the stability of shared items over time, using only common items from different measures cannot adequately capture the kinds of meaningful change in the expression of a phenomenon that justified the measurement change in the first place. For example, although "fighting with others" is considered an expression of externalizing at all ages, fighting during childhood has a different meaning than fighting during adulthood (Patterson, 1993); therefore, using only this item to assess externalizing behavior would not fully capture meaningful shifts in the expression of externalizing problems across development. Likewise, anhedonia is a core feature of depression at all ages, but it may manifest as a lack of interest in toys during early childhood, a global sense of boredom in adolescence, and a lack of interest in sex during adulthood (Weiss & Garber, 2003); therefore, using a single item to assess anhedonia may not fully capture meaningful shifts in the expression of depression across development. Indeed, the ideal invariance approach must account for both common and unique items to capture the developmental phenomenon of interest fully and over time. Therefore, the goal of this article was to illustrate the adaptation of traditional approaches to evaluate longitudinal measurement invariance in order to harmonize measures of a common construct across shifting measures of depressive symptoms using both common and unique items across five waves of data from early adolescence through early adulthood.

METHOD

Participants and Procedures

The sample used for these analyses included 392 youth (52% female) who participated in a dual-site longitudinal study that investigated the role of parents in their children's development from early adolescence to young adulthood in Phoenix, Arizona, and Riverside, California. The study targeted two-parent families who were of European or Mexican descent with an adolescent who was enrolled in seventh grade. All three participating family members were required to be from the same ethnic background, and families were recruited to include both intact families (i.e., two biological parents in the household) and stepfather families (i.e., a biological mother and a male romantic partner who was acting as a "father figure" to the child in the residence). The father and the mother were not required to be legally married, but the household structure had to be in place for more than 1 year.

The resulting sample included 110 European American intact families (96.36% married), 89 European American stepfather families (75.28% married), 107 Mexican American intact families (94.39% married), and 86 Mexican American stepfather families (44.19% married). Assessments began when the adolescents were enrolled in seventh grade (M_{age_W1} = 12.89, SD = .48) and lasted until they were young adults (M_{age_W5} = 21.95, SD = .77; N = 276), with intervening assessments at Wave 2 (M_{age_W2} = 13.89, SD = .76; N = 365), Wave 3 (M_{age_W3} = 15.53, SD = .65; N = 321), and Wave 4 (M_{age_W4} = 19.68, SD = .70; N = 287). The annual adjusted family income ranged from $8,000 to over $100,000, with a mean of $67,410.06 ($SD$ = $47,194.79), though 19.6% of the families earned below $35,000 per year. There was no significant difference in family income between intact (M = 66,705.17, SD = 47,151.39) and stepfather families (M = 68,362.45, SD = 47,489.87), $t(389) = .34$, *ns*. However, European American families reported higher household income (M = 86,678.08, SD = 54,392.10) than Mexican American families (M = 47,514.62, SD = 26,588.13), $t(289.79) = 9.09$, $p < .001$. Across the five data waves, 377 (96.2%) of the families completed two or more assessments. The 15 youth who did not participate in two or more assessments reported marginally higher rates of depressive symptoms at Wave 1 than youth who returned for one or more follow-ups, $t(14.563) = 2.089$, $p = .055$.

The recruitment procedures for this study varied by collection site because of different state laws and school district policies (see Stevenson et al., 2014, for description). Upon determining eligibility and acquiring consent from

each parent and assent from the adolescent, participants completed a full battery of assessments administered at the research site, during home visits, or via phone that lasted about three hours in their preferred language (English or Spanish). Assessments were conducted using the same procedures across all waves, and each family member received monetary compensation for their time. All procedures for this study were approved by the Institutional Review Boards of the participating universities.

Measures

Depressive Symptoms

Youth's depressive symptoms were assessed by self-reports. At Waves 1 through 3, depressive symptoms were assessed using eight items from the 27-item Child Depression Inventory (CDI; Kovacs, 1992). Sample items (e.g., in the past month, things bothered me) were scored on a 3-point scale, from 1 (e.g., things bothered me all the time) to 3 (e.g., things bothered me once in a while). Four items, including the sample item just listed, were reverse coded and then composited so that higher scores reflected higher levels of depressive symptoms (α = .652–.718). The CDI was abbreviated for use in this study because of time constraints. Employing data from the full CDI scale administered in prior work (Wolchik et al., 2000), stepwise regression analyses were used to identity the items that accounted for 90% of the variance in the full scale score (see Schenck et al., 2009, for a full description). At Waves 4 and 5, depressive symptoms were assessed using 11 items (e.g., I feel lonely) from the 18-item Anxious/Depressed subscale of the Adult Self Report (ASR; Achenbach & Rescorla, 2003) . Items were rated on a 3-point scale, from 1 (*not true*) to 3 (*very true or often true*), α = .792–.839. Both the CDI and ASR are well-established assessments of depressive symptoms and have been validated in different ethnic-racial populations using clinical and nonclinical samples (Rescorla & Achenbach, 2004; Ruggiero, Morris, Beidel, Scotti, & McLeer, 1999). Although the abbreviation of the CDI constrained our ability to describe clinical levels of depressive symptoms at earlier waves, Anxious/Depressed subscale scores on the ASR in late adolescence indicated that approximately 12% of the participants evidenced borderline elevations in symptomatology (i.e., *T* score ≥ 65) and 5% evidenced clinical levels of symptomatology (i.e., *T* score > 69; Achenbach & Rescorla, 2003).

Data Preparation and Analytic Plan

Prior to conducting longitudinal measurement invariance analyses for depressive symptoms, we employed a parceling technique wherein measurement items were compared across all waves to identify items that assessed the same symptom across waves (i.e., common items) and those that assessed varied symptoms across measurement waves (i.e., unique items; Kishton & Widaman, 1994). Common items were identified using a content validation approach (Haynes, Richard, & Kubany, 1995) based on theoretical and conceptual overlap, as well as agreement across the first two authors. Standard approaches to parceling require that items conform to a unidimensional scale. Depressive symptoms were assessed at Waves 1, 2, and 3 with a depression scale that exhibited unidimensionality in prior research. In contrast, the 18-item Anxious/Depressed subscale of the ASR was used at Waves 4 and 5, so we conducted an exploratory factor analysis to extract Anxiety and Depression factors from this instrument. Findings from a two-factor solution using the data from Wave 4 revealed 11 items that loaded on a unidimensional depression symptom factor and six items that loaded on an anxiety symptom factor, root mean square error of approximation (RMSEA) = .056, 90% confidence interval (CI) [.044, .067]. One item (i.e., I lack self-confidence) was excluded from subsequent analyses because it cross-loaded on both factors. Given that a second independent sample was not available for these analyses, we conducted a confirmatory factor analysis using data from Wave 5, which supported this two-factor structure, RMSEA = .056, 90% CI [.044, .067].

Depressive items were distributed into four parcels at each wave of measurement (i.e., four indicators for each latent variable) to ensure model identification and convergence (Kline, 2015). Table 1 depicts each depressive item and its corresponding parcel designation. Common items across all waves were summed to create two unidimensional parcels and to reduce unwanted error variance in the data (Kishton & Widaman, 1994; Little, Cunningham, Shahar, & Widaman, 2002). The remaining unique items at each wave were used to create a second set of unidimensional parcels for each wave. Thus, depressive symptoms were assessed with four parcels at each wave of measurement, corresponding to two "common item" parcels that were constructed in the same fashion across all five waves of measurement and two "unique item" parcels for each wave of measurement. The unique item parcels were composited in identical fashion across Waves 1, 2, and 3, and the other set of unique item parcels were constructed in identical fashion across Waves 4 and 5. The correlations, means, and standard deviations for the common and unique parcels of youth's depressive symptoms are shown in Table 2.

Factorial invariance analyses were conducted in Mplus 7.4 (Muthén & Muthén, 1998–2012) to evaluate how well successive invariance models fit the data for depressive symptoms (Widaman et al., 2010). Missing data were handled using full-information maximum likelihood estimation (Arbuckle, 1996). Three sets of factorial invariance models were computed. First, factorial invariance models were computed with data from the first three waves of assessment given that the same items were administered at these data points. Second, factorial invariance models were computed with data from the final two waves of

TABLE 1
Common and Unique Items for Youth's Depressive Symptoms

CDI (Waves 1–3)	*ASR (Waves 4–5)*	*Type of Item*	*Parcel*
2. Could not make up my mind about things	13. I feel confused or in a fog	Common	com1
5. Think about killing myself	91. I think about killing myself	Common	com1
6. Feel alone	12. I feel lonely	Common	com2
8. As good as other kids	35. I feel worthless or inferior	Common	com2
1. Things bothered me		Unique	uniq1
3. My looks		Unique	uniq1
4. I had trouble sleeping		Unique	uniq2
7. My school work		Unique	uniq2
	14. I cry a lot	Unique	uniq1
	31. I am afraid I might think or do something bad	Unique	uniq1
	33. I feel that no one loves me	Unique	uniq1
	34. I feel that others are out to get me	Unique	uniq2
	52. I feel too guilty	Unique	uniq2
	103. I am unhappy, sad, or depressed	Unique	uniq1
	107. I feel that I can't succeed	Unique	uniq2

Note: CDI = Child Depression Inventory; ASR = Adult Self Report; com = Common Parcel of Depressive Symptoms; uniq = Unique Parcel of Depressive Symptoms.

TABLE 2
Correlations, Means, and Standard Deviations for the Common and Unique Parcels of Youth's Depressive Symptoms

Variables	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
1. w1com1	—																			
2. w1com2	.32	—																		
3. w1uniq1	.36	.45	—																	
4. w1uniq2	.20	.36	.35	—																
5. w2com1	.27	.22	.28	.27	—															
6. w2com2	.16	.31	.27	.23	.37	—														
7. w2uniq1	.14	.22	.30	.18	.31	.50	—													
8. w2uniq2	.20	.19	.24	.31	.44	.47	.38	—												
9. w3com1	.15	.18	.24	.21	.25	.19	.22	.22	—											
10. w3com2	.20	.30	.32	.10	.17	.30	.30	.22	.44	—										
11. w3uniq1	.07	.19	.26	.13	.25	.27	.38	.28	.36	.43	—									
12. w3uniq2	.17	.28	.21	.26	.22	.18	.14	.37	.32	.36	.25	—								
13. w4com1	.00	.12	.11	.08	.16	.11	.10	.12	.23	.16	.26	.13	—							
14. w4com2	.00	.16	.24	.14	.16	.07	.13	.04	.21	.19	.34	.15	.57	—						
15. w4uniq1	.02	.20	.18	.13	.18	.11	.10	.16	.26	.18	.29	.21	.64	.69	—					
16. w4uniq2	−.06	.16	.19	.19	.13	.15	.07	.13	.17	.18	.25	.23	.50	.56	.67	—				
17. w5com1	.07	.08	.15	.08	.13	.11	.19	.21	.19	.19	.31	.13	.31	.24	.31	.30	—			
18. w5com2	.04	.13	.18	.12	.10	.17	.23	.14	.17	.17	.35	.09	.33	.45	.39	.36	.57	—		
19. w5uniq1	.04	.19	.16	.16	.10	.08	.14	.10	.23	.19	.22	.15	.43	.44	.60	.45	.51	.61	—	
20. w5uniq2	.05	.11	.14	.16	.11	.12	.13	.15	.14	.11	.20	.13	.33	.34	.39	.43	.40	.56	.53	—
M	1.44	1.42	1.42	1.41	1.39	1.40	1.35	1.42	1.42	1.41	1.42	1.51	1.21	1.29	1.22	1.18	1.25	1.32	1.19	1.16
SD	.38	.49	.44	.50	.39	.44	.44	.47	.34	.44	.43	.50	.33	.40	.33	.29	.33	.40	.28	.27

Note: com = Common Parcel of Depressive Symptoms; uniq = Unique Parcel of Depressive Symptoms.

assessment. Third, a final series of factorial invariance models combined data across all five waves of assessment.

Satorra's (2000) likelihood ratio chi-square difference test evaluated comparative fit across each pair of nested models. However, the likelihood ratio test is influenced by sample size (Browne & Cudeck, 1993), such that, when sample size is large (as in the current application), differences in model fit can be deemed statistically significant prompting a rejection of the constraints invoked even if the differences in fit are of trivial magnitude. Therefore, we also examined practical fit indices that are relatively unaffected by sample size, including the Tucker–Lewis index (TLI; Tucker & Lewis, 1973), comparative fit index (CFI; Bentler, 1990), RMSEA

(MacCallum, Browne, & Sugawara, 1996), and standardized root mean square residual (SRMR; Hu & Bentler, 1999). *Close* fit of a model to data is indicated by TLI and CFI values greater than .95 (Bentler, 1990; Tucker & Lewis, 1973), RMSEA values less than .05 (Browne & Cudeck, 1993; but see Hu & Bentler, 1999, who suggested a value of less than .06), and SRMR values below .08 (Hu & Bentler, 1999). The RMSEA is accompanied by a 90% CI; if the lower limit of the CI falls below .05, close fit of the model to the data cannot be rejected (MacCallum et al., 1996). Ideally, all practical fit indices for a model would indicate close fit of the model to data, and changes in fit indices across models of less than about .01 are typically deemed relatively unimportant. If one or more practical fit indices is not in the range of close fit, or if relatively large changes in practical fit occur when invoking constraints, overall model fit may be considered unacceptable and the model rejected; however, minor modifications can often restore close model fit.

Model respecifications can be aided by modification indices. A modification index (MI) is an estimate of the change in the model chi-square that would accompany freeing a fixed or constrained parameter estimate. Because each MI is associated with 1 degree of freedom, an MI value of 3.84 or larger indicates that a significant improvement in model fit at the .05 level would occur if the parameter were freely estimated. MIs should be used to improve model fit to acceptable levels only if they are theoretically justified and greater than 3.84.

TABLE 3
Fit Indices for the Factorial Invariance Models for Youth's Depressive Symptoms

Model	χ^2	df	*RMSEA [CI]*	*CFI*	*TLI*	*SRMR*
Waves 1–3						
Model 1A	112.14	51	.055 [.041, .069]	.934	.915	.045
Model 1B	62.28	43	.034 [.011, .051]	.979	.968	.035
Model 1C	69.97	49	.033 [.012, .050]	.977	.970	.044
Model 1D	82.74	55	.036 [.018, .051]	.970	.964	.049
Model 1E	115.75	63	.046 [.033, .059]	.943	.941	.068
Waves 4–5						
Model 2A	66.94	19	.090 [.067, .113]	.953	.930	.039
Model 2B	23.76	15	.043 [.000, .074]	.991	.984	.034
Model 2C	36.49	18	.057 [.030, .084]	.982	.972	.056
Model 2D	52.44	21	.069 [.046, .093]	.969	.959	.048
Model 2E	55.44	25	.062 [.040, .084]	.970	.966	.055
All Waves						
Model 3A	275.82	160	.043 [.034, .051]	.942	.931	.047
Model 3B	181.76	146	.025 [.010, .036]	.982	.977	.043
Model 3C	204.51	156	.028 [.016, .038]	.976	.970	.052
Model 3D	266.01	166	.039 [.030, .048]	.950	.942	.056
Model 3E	236.94	165	.033 [.023, .043]	.964	.958	.054
Model 3F	334.20	179	.047 [.039, .055]	.922	.917	.076

Note: RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual.

RESULTS

Factorial Invariance Models for Waves 1–3

A baseline configural invariance model (Model 1A) evaluated whether the same pattern of fixed and free loadings characterized youth's depressive symptoms across the first three waves of assessment. The mean of the latent variable at Wave 1 was fixed at zero and the variance was fixed at 1, whereas the means and variances for the latent variables for the two remaining waves were freely estimated. All factor loadings, intercepts, and variances of the common and unique parcels at each wave were freely estimated. Results from Model 1A revealed that, as expected, the statistical test of fit of the model to the data was significant, $\chi^2(51) = 112.14$, $p < .001$, providing a statistical basis for rejecting the model. Practical fit indices were more mixed, as the RMSEA at .055, 90% CI [.041, .069], and the SRMR at .045 both implied close model fit, but the CFI of .934 and TLI of .915 reflected less-than-close model fit (see Table 3).

Longitudinal models often require inclusion of across-wave covariances between unique factors for identical indicators across times of measurement, reflecting reliable variance in indicators that is unrelated to the latent variable but consistent across time. Therefore, a second configural invariance model (Model 1B) was specified to include all across-wave, first-order unique factor covariances (e.g., the unique factor for the first common parcel of depression at Wave 1 was allowed to covary with the unique factor for the first common parcel of depressive symptoms at Wave 2). Results from Model 1B showed substantially improved fit, with $\Delta\chi^2(8) = 49.86$, $p < .001$. Furthermore, although the overall index of statistical model fit remained significant, $\chi^2(43) = 62.28$, $p = .029$, all of the practical fit indices revealed close fit of the Model 1B to the data, with RMSEA = .034, 90% CI [.011, .051], CFI = .979, TLI = .968, and SRMR = .035.

Following the identification of Model 1B as the best fitting configural invariance model, a weak factorial invariance model (Model 1C) tested whether factor loadings were invariant across time. Given that the parcels were formed using the same items across the three waves, the factor loadings for similar parcels were constrained to be equal. Model 1C evidenced a nonsignificant drop in fit when compared to Model 1B, $\Delta\chi^2(6) = 7.69$, $p = .261$. Moreover, all practical fit indices indicated that Model 1C fitted the data closely, with RMSEA = .033, 90% CI [.012, .050], CFI = .977, TLI = .970, and SRMR = .044. Results from this model indicated that the latent variables of depressive symptoms assessed the same underlying construct across time.

Building on the weak factorial invariance model, a strong factorial invariance model (Model 1D) evaluated whether the intercepts of the parcels were equivalent across time. Thus, the intercepts for each item parcel were

constrained to be equal across waves. Although this led to a statistically significant drop in fit when compared to Model 1C, $\Delta\chi^2(6) = 12.77$, $p = .047$, the strong invariance model continued to evidence close fit to the data; the practical fit indices showed little change from those for Model 1C, with RMSEA = .036, 90% CI [.018, .051], CFI = .970, TLI = .964, and SRMR = .049. The close fit of the more parsimonious Model 1D suggests that the intercepts for the parcels of depressive symptoms can be constrained to be equal across time with little harm to the fit of the model to the data.

Finally, a strict factorial invariance model (Model 1E) evaluated whether the unique variances for the parcels were equal across time. Equality constraints were added to the unique factor variances of all the corresponding item parcels across time. Model 1E evidenced a relatively large and statistically significant drop in fit, $\Delta\chi^2(8) = 33.01$, $p < .001$, relative to Model 1D. Furthermore, although the RMSEA = .046, 90% CI [.033, .059], and SRMR = .068 implied close fit, the CFI = .943 and TLI = .941 exhibited rather large decreases and fell below the standard of .95 for close model fit. Therefore, we selected the strong factorial invariance model (Model 1D; see Figure 1), as the best-fitting model for the items assessing youth's depressive symptoms across the first three waves of assessment.

Factorial Invariance Models for Waves 4 and 5

A second set of factorial invariance models were computed using the items administered at Waves 4 and 5. Similar to the earlier waves of data, a baseline configural invariance model (Model 2A) evaluated whether the same pattern of relations was observed in youth's depressive symptoms across the final two waves of assessment. The mean of the first latent variable was fixed at zero and the variance was fixed at 1. In addition, the factor loadings, intercepts, and unique variances of the common and unique parcels at both waves were freely estimated. Results from Model 2A suggested that the model fit to the data was poor with regard to statistical fit, $\chi^2(19) = 66.94$, $p < .001$, and with regard to several measures of practical fit, with RMSEA = .090, 90% CI [.067, .113], CFI = .953, TLI = .930, and SRMR = .039 (see Table 3). Consistent with the preceding modeling of data from Waves 1–3, a second configural invariance model (Model 2B) was specified to include all across-wave, first-order covariances among unique factors for the same parcel across time. Model 2B evidenced a large and significant improvement in fit, $\Delta\chi^2(4) = 43.19$, $p < .001$, and suggested a close fit to the data for all practical fit indices, with RMSEA = .043, 90% CI [.000, .074], CFI = .991, TLI = .984, and SRMR = .034.

Building on Model 2B, a weak factorial invariance model (Model 2C) that constrained factor loadings to be equal across time was estimated. Although Model 2C evidenced a statistically significant drop in fit when compared to Model 2B, $\Delta\chi^2(3) = 12.74$, $p = .005$, the fit indices for the overall

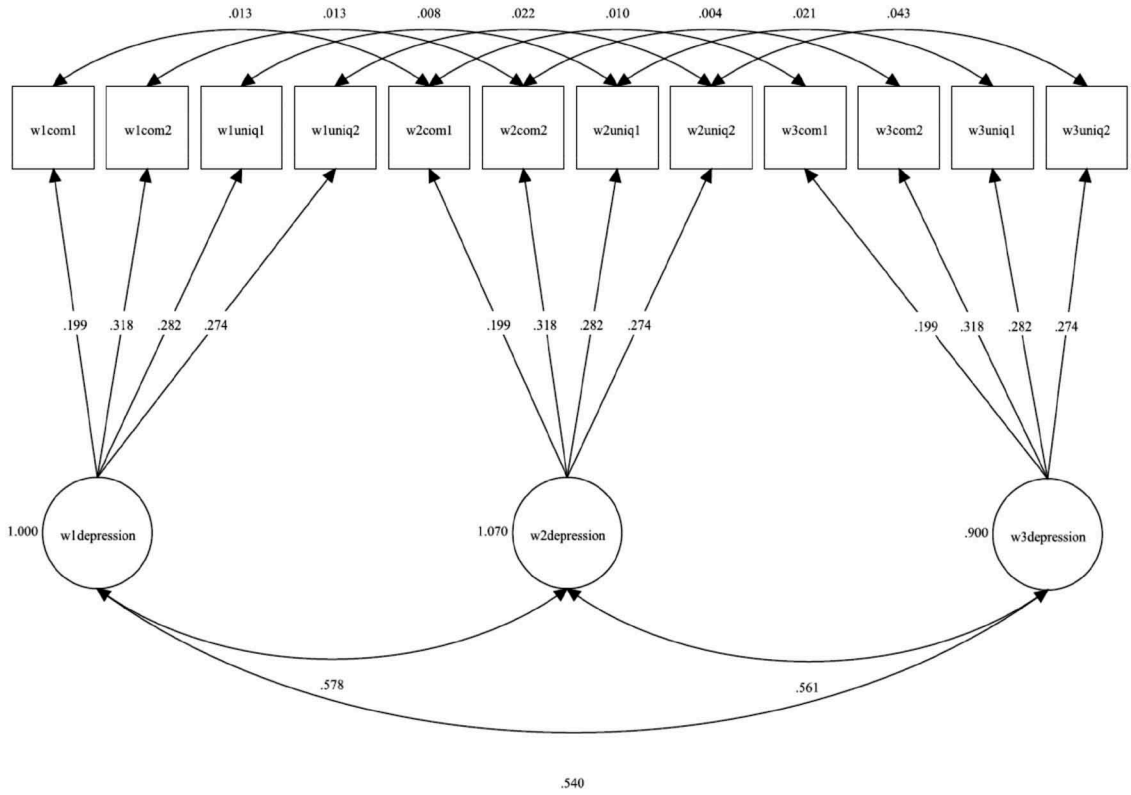

FIGURE 1 Strong factorial invariance model for youth's depressive symptoms with unstandardized parameter coefficients for Waves 1–3.

model indicated that Model 2C fit the data closely, with RMSEA = .057, 90% CI [.030, .084], CFI = .982, TLI = .972, and SRMR = .056. Therefore, a strong factorial invariance model (Model 2D) that included equality constraints for the intercepts of the item parcels was evaluated. This model also evidenced a statistically significant drop in fit, $\Delta\chi^2(3) = 15.95$, $p = .001$. However, three of the practical fit indices were still solidly in the range of close fit, with CFI = .969, TLI = .959, and SRMR = .048. Further, although the point estimate of the RMSEA fell above recommended cutoffs, with RMSEA = .069, 90% CI [.046, .093], the lower limit of the CI fell below .05, so close fit could not be rejected (cf. MacCallum et al., 1996). Finally, a strict factorial invariance model (Model 2E) that invoked equality constraints on the unique factor variances of the item parcels was evaluated. Compared to Model 2D, Model 2E did not show a significant drop in fit, $\Delta\chi^2(4) = 3.00$, $p = .558$, and fit the data somewhat more closely than did Model 2D, with RMSEA = .062, 90% CI [.040, .084], CFI = .970, TLI = .966, and SRMR = .055. Therefore, the more parsimonious strict factorial invariance model (Model 2E, see Figure 2) was retained as the best-fitting model for the items assessing youth's depressive symptoms for the final two waves of assessment.

Factorial Invariance Models Using All Waves of Data

A final set of factorial invariance models were fitted to the data across all five waves assessing youth's depressive symptoms. First, a baseline configural invariance model (Model 3A) evaluated whether the same pattern of fixed and free loadings was evident for depressive symptoms across time. Similar to the previous models, the mean for the first latent variable was fixed at zero and the variance was fixed at one. The first common parcel of depressive symptoms served as the anchor indicator for the baseline configural invariance model. For this parcel, the factor loading and the intercept were fixed to be equal across all five waves of data, whereas the remaining factor loadings, intercepts, and unique variances were freely estimated. In addition, the means and variances for the latent variables at the last four waves of measurement were freely estimated. Results from Model 3A revealed fit to the data that was not fully acceptable, with $\chi^2(160) = 275.82$, $p < .001$, CFI = .942, and TLI = .931, even though the RMSEA = .043, 90% CI [.034, .051] and SRMR = .041 met cutoffs for close fit. Consistent with previous models, first-order covariances among unique factors across waves were added to produce a second configural invariance model (Model 3B). Note that, in this model, unique factor covariances were allowed for the two common parcels across all adjacent waves of measurement (e.g., from Wave 1

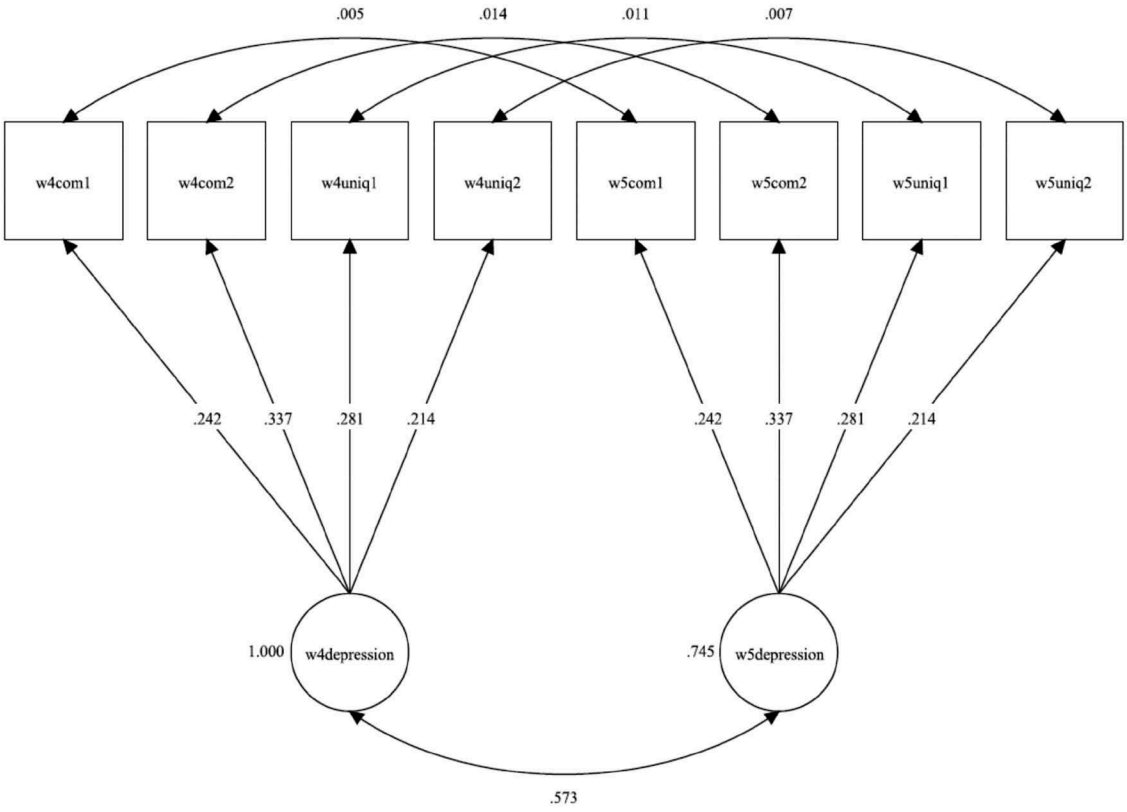

FIGURE 2 Strict factorial invariance model for youth's depressive symptoms with unstandardized parameter coefficients for Waves 4–5.

to Wave 2, Wave 2 to Wave 3) yielding eight estimates. Next, unique factor covariances for the unique indicators were allowed from Wave 1 to Wave 2, Wave 2 to Wave 3, and Wave 4 to Wave 5 (but not from Wave 3 to Wave 4, as these parcels were composed of different items), for an additional six estimates. The results for Model 3B yielded a significant statistical index of fit, $\chi^2(146) = 181.76$, $p = .024$, but all practical fit indices indicated close fit of the model, with RMSEA = .025, 90% CI [.010, .036], CFI = .982, TLI = .977, and SRMR = .043, and Model 3B fit substantially better than Model 3A statistically, $\Delta\chi^2(14) = 94.06$, $p < .001$.

Following identification of Model 3B as the best-fitting configural factorial invariance model, a weak factorial invariance model (Model 3C) tested whether the factor loadings for depressive symptoms were invariant across time. At each assessment wave, the loadings for the two common item parcels were fixed to be equal across the five data points. Given that the unique parcels for depressive symptoms were created using the same measures at the first three waves, the factor loadings for these unique parcels were constrained to be equal across these waves, whereas the unique parcels for depressive symptoms at Waves 4 and 5 were constrained to be equal to one another. Although Model 3C evidenced a statistically significant drop in fit to the data relative to Model 3B, $\Delta\chi^2(10) = 22.75$, $p = .012$, all practical fit indices exhibited little change from those for Model 3B and indicated Model 3C fit the data closely, with RMSEA = .028, 90% CI [.016, .038], CFI = .976, TLI = .970, and SRMR = .052. Results from this model indicated that the latent variables of depressive symptoms assessed the same underlying construct across time with standardized factor loadings for the item parcels across waves ranging from .52 to .74.

A strong factorial invariance model (Model 3D) evaluated whether the intercepts of the common and unique parcels for youth's depressive symptoms were invariant across time. Equality constraints were applied to intercepts for item parcels with the same items across waves. For example, the intercepts for the common parcels were constrained to be equal across all five waves of measurement. At Waves 1–3, the intercepts for unique parcels were constrained to be equal, and similar equality constraints were added to the unique parcels at Waves 4–5. This strong invariance model revealed a fit to the data that was not fully acceptable, $\chi^2(166) = 266.01$, $p < .001$, which was substantially worse than Model 3C, $\Delta\chi^2(10) = 61.50$, $p < .001$, and most practical fit indices were substantially reduced relative to Model 3C, with RMSEA = .039, 90% CI [.030, .048], CFI = .950, and TLI = .942. Therefore, based on a modification index, we formulated Model 3E by relaxing the intercept constraint for the first common parcel at Wave 4, which improved the fit of the model substantially over that of Model 3D, $\Delta\chi^2(1) = 29.07$, $p < .001$. Furthermore, all practical fit indices returned to levels that indicated close fit of the model to the data, with RMSEA = .033, 90% CI [.023, .043], CFI = .964, TLI = .958, and SRMR = .054.

Building on Model 3E, a strict factorial invariance (Model 3F) invoked equality constraints on the unique factor variances for parcels with similar items across waves. All the unique factor variances for the common parcels were constrained to be equal across time, whereas the unique factor variances for the unique parcels at Waves 1–3 and for the unique parcels at Waves 4–5 were constrained to be equal. The overall fit of the model was poor, $\chi^2(179) = 334.20$, $p < .001$, and significantly worse than Model 3E, $\Delta\chi^2(14) = 97.26$, $p < .001$. In addition, all practical fit indices exhibited a clear decline in fit from Model 3E, with RMSEA = .047, 90% CI [.039, .055], CFI = .922, TLI = .917, and SRMR = .076. Therefore, the partial strong factorial invariance model (Model 3E; See Figure 3) was retained as the best-fitting model for these data. In this model, full, appropriate invariance of all factor loadings was evident, such that the factor loadings for the two common parcels were invariant across all five waves of measurement, the factor loadings for the two unique parcels at Waves 1–3 were invariant across these three waves, and the factor loadings for the two unique parcels at Waves 4 and 5 were also invariant across time. In addition, all parcel intercepts evidenced the same pattern of invariance across time with a single exception, the noninvariant intercept for the first common parcel at Wave 4, when the measurement of the construct changed. Invariance constraints could not be enforced on unique factor variances across time without a significant decline in model fit and an overall poor fit to the data, so strict measurement invariance was not tenable. However, given the fit of the partial strong invariance model, the model provided a sufficient and justifiable basis for investigating differences in mean and variance on the latent construct of depressive symptoms across all five waves of measurement.

DISCUSSION

This investigation illustrated how traditional approaches to evaluating measurement invariance can be adapted to assess longitudinal invariance in psychological constructs, even when measure instruments change across multiple developmental periods. Using both common and unique items assessing youth's depressive symptoms from early adolescence through young adulthood, this study was able to establish at least strong factorial invariance in both measures of depression. Furthermore, we were able to establish partial strong factorial invariance when analytic models were fitted to the data across all five waves of measurement.

These data demonstrate that a sequential analytic approach that begins with the parceling of common items

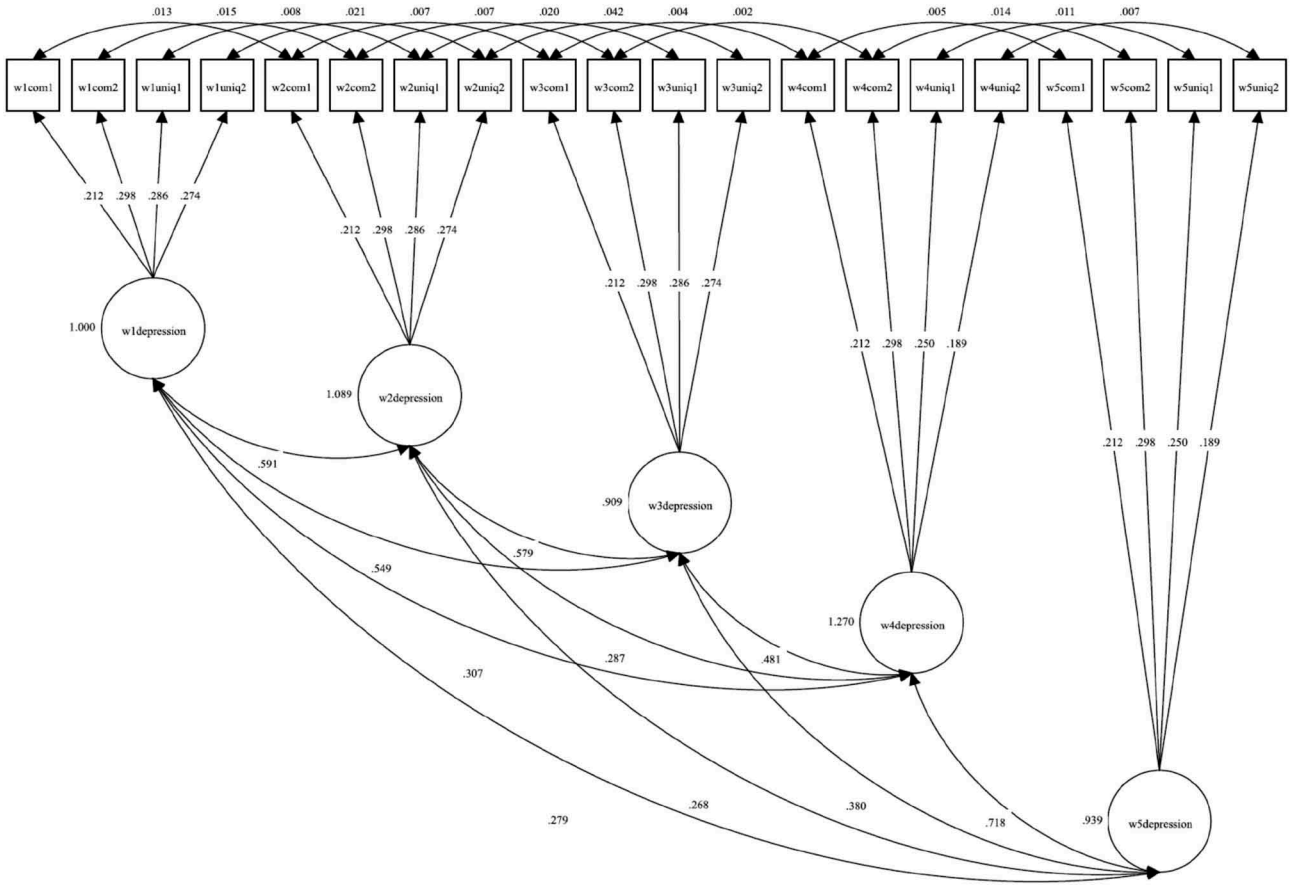

FIGURE 3 Partial strong factorial invariance model for youth's depressive symptoms with unstandardized parameter coefficients for Waves 1–5.

across measurement occasions and unique items within occasions can be used to determine whether the same underlying construct is being assessed across different measurements across time. The capacity to demonstrate longitudinal measurement invariance across different assessments will enable developmental scientists to examine both intraindividual and interindividual differences in complex psychological constructs across time. Further, the parceling and linking of items across time can be extended to support researchers' efforts to investigate both trait-invariant and time-varying cross-lagged panel models (Berry & Willoughby, 2016; Cole et al., 2017; Hamaker et al., 2015). Although the primary contribution of this methodological approach is to assist researchers in understanding true changes in depressive symptoms across time, the information gained from using this approach will allow clinicians to understand more about the emergence, stability, and pattern of depressive symptoms over adolescence in different populations and treatment contexts.

Two fundamental features of this data analytic approach were the use of common items and item parceling. Similar to IRT, common items were used as anchor indicators to link the measurement scales at the latent level across the different measures of depressive symptoms. This demonstrates that common items are essential for any analytic approach that seeks to establish longitudinal measurement invariance using different measurements because at least some items must allow for linking on a common scale. However, whereas IRT involves a computationally intensive process to establish measurement invariance, the current analytical approach may be a preferred method for researchers who are more familiar with traditional factorial invariance models and seek a more accessible approach for modeling developmental change over time when change in measurement is required. Given the novelty of this approach, it is unclear as to how many common items are required to establish longitudinal measurement invariance. For these data, we used two items to create each common parcel, and a minimum of two common parcels was essential for testing invariance of the linking of latent scores on a common scale. Similarly, we used a content validation approach to determine the commonality of each item across time. As a result, common items were chosen based on theoretical and conceptual similarity and consensus between two content evaluators. However, future research should consider additional alternatives, such as the

utilization of inter-item correlations as well as examining whether associations between common items and external criteria are similar or invariant across waves.

Despite advances in assessing invariance across changing measurement, it is important to acknowledge that the measures in this study did not vary in method (i.e., both measures were self-report surveys). Further research is needed to evaluate whether this approach will work for investigations in which there are shifts in measuring instruments as well as methods (e.g., from observations in infancy to self-report data in adolescence) and/or informants (e.g., from parents or teachers to examiners or participants). One major concern is that informants provide information that might be context dependent (Kraemer et al., 2003), which can produce additional variation in scores that may require the use of additional analytic strategies. Recent work by Cole et al. (2017) evaluating time-invariant and time-varying differences in youth's depressive symptoms found congruence across multiple informants (e.g., self, parents, peers, teachers), with some reporters showing more congruence than others in their assessment of trait-invariant versus time-varying dimensions of depression. Therefore, it is plausible that the current approach could be used to evaluate longitudinal measurement invariance when multiple, but appropriately congruent, informants are used.

In addition to the use of common items, the current approach to establishing longitudinal measurement invariance used item parcels to reduce unwanted variation in the data because of the large number of items and different measures in the study. Although this risk is reduced when item parcels are unidimensional, which was the case in this study, some scholars argue against the use of parcels under any circumstance because they can mask misspecification issues and produce misleading fit indices (e.g., Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013). However, among others, Little, Rhemtulla, Gibson, and Schoemann (2013) contended that parceling does not lead necessarily to biased or confounded results when used judiciously, and recent research by Cole, Perkins, and Zelkowitz (2016) and by Rhemtulla (2016) suggests that careful parceling can lead to optimal analytic outcomes. Future research could compare results using our proposed analytic strategy at the item level with results at the parcel level to evaluate whether item parceling has any discernable effect on results. We opted to use item parcels to keep the number of indicators to a manageable level and to employ indicators that have more optimal psychometric properties in our models. If future research demonstrates that the use of item parcels introduces substantial levels of bias, then our advocating for the use of item parcels should be qualified.

Another area of contention is the use of chi-square values, practical fit indices, and MIs to make decisions about model fit. Recently, scholars have advocated for the use of multiple fit indices to evaluate model fit rather than using absolute cutoffs for single indices (Chen, Curran, Bollen, Kirby, & Paxton, 2008). Although we used the chi-square value to compare nested models, we adopted a holistic approach to make decisions about model fit, which included consideration of multiple practical fit indices. Consideration of agreement across multiple practical fit indices can ensure a more balanced evaluation of model fit than reliance on only a single index. Relatedly, in one instance, an MI was used to improve model fit to acceptable levels. Although MIs are exploratory and atheoretical when generated, scholars have argued that model modifications suggested by MIs can be applied justifiably if a researcher deems them to be supported by extant theory or reasonable a priori interpretation (Whittaker, 2012). Considering that a number of parcels in our study were created with identical or conceptually similar items, but others were not, the inclusion of across-wave covariances between unique factors for "common item" indicators was required across all waves to account for variance that was consistent across waves yet unrelated to the latent variables. However, we would never have allowed covariances between "unique item" indicators at Waves 3 and 4—even in the presence of large MIs for these estimates—because the resulting covariances would have had no conceptual interpretation, given the different content in these "unique item" parcels across measures. Furthermore, MIs led to model respecifications only if their inclusion lead to significant improvement in the model chi-square as indicated by an MI of 3.84 or larger, where 3.84 is the critical value of chi-square with 1 *df* at the $\alpha = .05$ level. Despite these cautionary measures, some scholars suggest that MIs are trustworthy only when used in conjunction with expected parameter change values, when sample size is greater than 100 observations, when standardized factor loadings are higher than .40, and when factor interrelations are greater than .20 (Whittaker, 2012). All our initial configural models met these criteria, but we continue to stress theory and interpretation as more important criteria when employing MIs to guide model respecification.

The contributions of this study to the extant literature on depression across adolescence and to clinical research methods should be considered in light of several limitations. First, the current investigation included a sample of Mexican and European American youth living in the southwestern United States. Therefore, these findings might not reflect the experiences of youth from the broader population. Further, in the absence of sufficient data to characterize the degree to which levels of depressive symptoms in this community sample reached clinical levels, caution is warranted when generalizing to other clinical and nonclinical samples. Second, these analyses did not test for potential differences in the expression of depressive symptoms as function of participants' sex, ethnicity-race, and socioeconomic status as these were not of focal interest. In a recent study using this same data set, we observed significant differences in internalizing symptoms by youth sex, but not by ethnicity-race (Tyrell, Yates, Reynolds, Fabricius, & Braver, 2018). Third, we conducted exploratory

and confirmatory factor analyses of the ASR anxiety/depression scale using two different waves of data with the same participants to identify and confirm the depressive symptoms factor. Given this unconventional approach, future work is needed to confirm this factor structure in an independent sample. Finally, this study assessed depressive symptoms using the CDI and ASR. However, it may have been more beneficial to use the Youth Self-Report (Achenbach, 1991) and ASR given their conceptual and methodological overlap. Researchers who plan to examine psychological phenomena across multiple developmental periods should carefully consider their measurements in advance of data collection to ensure that assessments of the same psychological construct have overlapping items across waves.

Despite recent advances, developmental researchers still face serious challenges in ongoing efforts to accurately capture development across time, particularly in the context of heterotypic continuity. The current investigation introduced a new approach to evaluate longitudinal measurement invariance when measurement instruments change across time. Despite noted limitations, we hope that this study advances future developmental research and stimulate ongoing discussions about the complexities and opportunities presented by efforts to establish longitudinal measurement invariance across developmental time.

ACKNOWLEDGMENTS

This manuscript was completed as a partial requirement for a minor in quantitative methods. We are grateful to all the families who participated in this research project.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

This study was supported by the National Institute of Mental Health, under Grant MH64829 R01, the National Institute of Child Health and Human Development, under Grant RO1HD0566-06A1, and funding provided to the first author through a Ford Foundation Predoctoral Fellowship and a Postdoctoral Fellowship from the National Institute of Mental Health, under Grant T32 MH015755.

ORCID

Fanita A. Tyrell http://orcid.org/0000-0001-6419-6485

REFERENCES

Achenbach, T. M. (1991). *Manual for the youth self-report and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. (2003). *Manual for the ASEBA adult forms & profiles*. Burlington: University of Vermont, Research Center for Children, Youth, and Families.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–278). Mahwah, NJ: Lawrence Erlbaum Associates.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Berry, D., & Willoughby, M. T. (2016). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*(4), 1186–1206. doi:10.1111/cdev.12660

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494. doi:10.1177/0049124108314720

Cole, D. A., Martin, J. M., Jacquez, F. M., Tram, J. M., Zelkowitz, R., Nick, E. A., & Rights, J. D. (2017). Time-varying and time-invariant dimensions of depression in children and adolescents: Implications for cross-informant agreement. *Journal of Abnormal Psychology*, *126*(5), 635–651. doi:10.1037/abn0000267

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, *21*(2), 164–174. doi:10.1037/met0000047

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45. doi:10.1111/j.1745-3992.1991.tb00207.x

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology*, *44*(2), 365–380. doi:10.1037/0012-1649.44.2.365

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. doi:10.1037/a0038889

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247. doi:10.1037/1040-3590.7.3.238

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Khoo, S.-T., West, S. G., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301–317). Washington, DC: American Psychological Association.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement*, *54*(3), 757–765. doi:10.1177/0013164494054003022

Kline, R. B. (2015). *Principles and practice of structural equation modeling, fourth edition*. New York, NY: Guilford Publications.

Kovacs, M. (1992). *Children's depression inventory: Manual*. North Tonawanda, New York: Multi-Health Systems.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from

multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*(9), 1566–1577. doi:10.1176/appi.ajp.160.9.1566

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*(2), 151–173. doi:10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300. doi:10.1037/a0033266

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. doi:10.1037/1082-989X.1.2.130

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*(3), 257–284. doi:10.1037/a0032773

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. doi:10.1037/a0015857

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/bf02294825

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology*, *61*(6), 911–919. doi:10.1037/0022-006X.61.6.911

Rescorla, L. A., & Achenbach, T. M. (2004). The Achenbach System of Empirically Based Assessment (ASEBA) for ages 18 to 90 years. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults* (pp. 115–152). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, *21*(3), 348. doi:10.1037/met0000072

Ruggiero, K. J., Morris, T. L., Beidel, D. C., Scotti, J. R., & McLeer, S. V. (1999). Discriminant validity of self-reported anxiety and depression in children: Generalizability to clinic referred and ethnically diverse populations. *Assessment*, *6*(3), 259–267. doi:10.1177/107319119900600306

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). Boston, MA: Springer US.

Schenck, C. E., Braver, S. L., Wolchik, S. A., Saenz, D., Cookston, J. T., & Fabricius, W. V. (2009). Relations between mattering to step- and non-residential fathers and adolescent mental health. *Fathering*, *7*(1), 70–90. doi:10.3149/fth.0701.70

Sroufe, L. A., Egeland, B., & Carlson, E. A. (1999). One social world: The integrated development of parent-child and peer relationships. In W. A. Collins & B. Laursen (Eds.), *Relationships as developmental contexts. The Minnesota symposia on child psychology* (Vol. 30, pp. 241–261). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sroufe, L. A., & Jacobvitz, D. (1989). Diverging pathways, developmental transformations, multiple etiologies and the problem of continuity in development. *Human Development*, *32*(3–4), 196–203. doi:10.1159/000276468

Steinberg, L., & Thissen, D. (2013). Item response theory. In J. S. Commer & P. C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 336–373). New York, NY: Oxford University Press.

Stevenson, M. M., Fabricius, W. V., Cookston, J. T., Parke, R. D., Coltrane, S., Braver, S. L., & Saenz, D. S. (2014). Marital problems, maternal gatekeeping attitudes, and father-child relationships in adolescence. *Developmental Psychology*, *50*(4), 1208–1218. doi:10.1037/a0035327

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. doi:10.1007/bf02291170

Tyrell, F. A., Yates, T. M., Reynolds, C. A., Fabricius, W. V., & Braver, S. L. (2018). The unique effects of maternal and paternal depressive symptoms on youth's symptomatology: Moderation by family ethnicity, family structure, and child gender. *Development and Psychopathology*, 1–14. doi:10.1017/S0954579418000846

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology*, *15*(2), 403–430. doi:10.1017/S0954579403000221

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*(1), 26–44. doi:10.1080/00220973.2010.531299

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wolchik, S. A., West, S. G., Sandler, I. N., Tein, J.-Y., Coatsworth, D., Lengua, L., … Griffin, W. A. (2000). An experimental evaluation of theory-based mother and mother–Child programs for children of divorce. *Journal of Consulting and Clinical Psychology*, *68*(5), 843–856. doi:10.1037/0022-006X.68.5.843