Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Can Measurement Invariance Be Established in a Non-Ergodic Situation?

Y. Jo [†], and J.H. Ryoo [*†]
†Department of Education, Yonsei University, Seoul, 03722, Republic of Korea
*Corresponding author. Email: ryoox001@yonsei.ac.kr

Abstract

Many statistical analyses generalize group-level characteristics to individuals that are defensible only under ergodicity. Ergodicity is defined as circumstantial/structural conditions of stationarity over time and homogeneity across individuals, which informs that individuals inherit the group-level research findings. When these conditions are violated, interpretation at the individual-level from the group-level model would be contaminated due to the heterogeneity of the group. Importantly, checking the ergodicity assumption is critical to draw valid inferences, yet the assumption is often ignored or replaced by longitudinal measurement invariance (MI) testing that only evaluates the consistency of the measurement tool whether factor loadings and intercepts can be treated as comparable across occasions. This study aims to examine how non-ergodicity influences MI decisions using Monte Carlo simulations from a cross-classified factor model with person- and time-level random effects, while manipulating random-effect magnitude, mean and variance drift, and between-group shift in loading and intercepts. Results indicate design-dependence. In the ergodic benchmark, rejection rate of cross-sectional MIs stayed near 10%, but that of longitudinal MIs rose from approximately 20% (small random effects) to roughly 80% (large random effects) at both metric and scalar invariance steps. Mean drift mainly increased longitudinal scalar non-invariance (often > 50%), whereas variance drift had limited impact. On the other hand, between-group non-homogeneity produced clear rejections in the cross-sectional design. In sum, longitudinal MI failures should not be taken as definitive evidence of instrument change in case of considering ergodicity of heterogeneity and stationarity. By checking the ergodicity, researchers may benefit from reporting random-effect magnitudes and exploring random-parameter or approximate MI as a sensitivity analysis.

Keywords: Cross-classified Structure, DSEM, Ergodicity, Measurement Invariance, Simulation Study

1. Introduction

Statistical analyses often interpret group-level associations as describing within-person change, but such inference is justified only under ergodicity, where between-person statistics correspond to within-person statistics reflecting within-person dynamics (Hamaker, 2012; Molenaar, 2004). For example, in intensive longitudinal data on sleep hours and concentration, individuals who sleep less on average may report poorer average concentration, but an individual's slightly fewer number of sleeping hours does not always predict a noticeably worse day. An individual's concentration can vary across people and contexts, which indicates non-ergodicity. Ergodicity requires stationarity (stable distributional properties over time) and homogeneity (shared dynamic structure and parameters across individuals) in definition (Molenaar, 2004). When either one of the requirements fails, population summaries may not represent individual trajectories. Empirical studies inform that psychological processes frequently violate these requirements due to heterogeneity in developmental and temporal dynamics (Fisher et al., 2018).

There is an analogy between ergodicity and measurement invariance (MI), which examines whether an instrument functions equivalently across groups or time. MI of the instrument confirms that the latent values can be compared and thus, MI is often treated as a prerequisite for valid comparisons of latent values (van de Schoot et al., 2012; Widaman & Reise, 1997). On the other than, in longitudinal studies, both ergodicity and MI would be connected.

For example, these distinct tests effectively presume interpretability in individual-level and stability of the measurement model across repeated assessments. However, such connection would work negatively in case that the latent variable model is changing over time. Accordingly, MI failures may reflect non-stationary or heterogeneity in the data-generating or sampling process. Therefore, the results should not be interpreted as definitive evidence in the case that the instrument itself is flawed (Hamaker, 2012; Molenaar, 1999).

Although the literatures of non-ergodicity and measurement variance have largely evolved separately, impact of one on the other has not been reported in the literature. This study aims to examine how MI testing results behave under explicitly non-ergodic data and what different MI rejection patterns indicate related to the violations on ergodicity such as instability and non-homogeneity.

We reveal the relationships between ergodicity and MI using Monte Carlo simulations that embed MI testing within a controlled non-ergodic data-generating framework. We generate intensive longitudinal responses from a four-indicator single-factor model and manipulate random-effect magnitude in measurement parameters, non-stationarity in the latent distribution, and non-homogeneity via between-group differences in fixed measurement parameters. MI is evaluated in a two-wave, two-group CFA that compares early versus late occasions, focusing on equality constraints for factor loadings and item intercepts. We then compare metric and scalar rejection rates across ergodic and non-ergodic scenarios to support theory-aligned interpretations of MI in longitudinal research.

2. Method

2.1 Framework

Simulation was conducted within a cross-classified longitudinal measurement model in which item responses are indexed by persons and measurement occasions, allowing loadings and intercepts to vary across both dimensions (Asparouhov & Muthén, 2015; McNeish et al., 2021). For person i at time t, four indicators ($j = 1, \dots, 4$) followed

$$y_{jit} = v_{jit} + \lambda_{jit}\eta_{it} + e_{jit}, \quad e_{jit} \sim N(0, \theta_j)$$

with θ_j calibrated to achieve the scale reliability (McNeish et al., 2021). Measurement parameters were decomposed into fixed effects and cross-classified random effects,

$$\lambda_{jit} = \alpha_{\lambda j} + \pi_{\lambda ji} + \tau_{\lambda jt}, \quad v_{jit} = \alpha_{vj} + \pi_{vji} + \tau_{vjt},$$

where αs denote grand means, πs indicate person-level deviations with variance, ω, and τs indicate time-level deviations with variance, ψ, (Asparouhov & Muthén, 2015). Latent dynamics was modeled as an AR(1) process,

$$\eta_{it} = \mu_t + \phi\left(\eta_{i,t-1} - \mu_{t-1}\right) + \zeta_{it}, \quad \phi = 0.8, \zeta_{it} \sim \mathcal{N}(0, \sigma_t^2).$$

In this simulation, stationarity held when μ_t and σ_t^2 were constant over time. Non-homogeneity was introduced via group-specific fixed measurement parameters $\alpha_{\lambda jg}$ and/or α_{vjg} (Ryu, 2014). Figure 1 summarizes the measurement model (Panel a) and the time- and person-level random-effect components (Panels b and c).

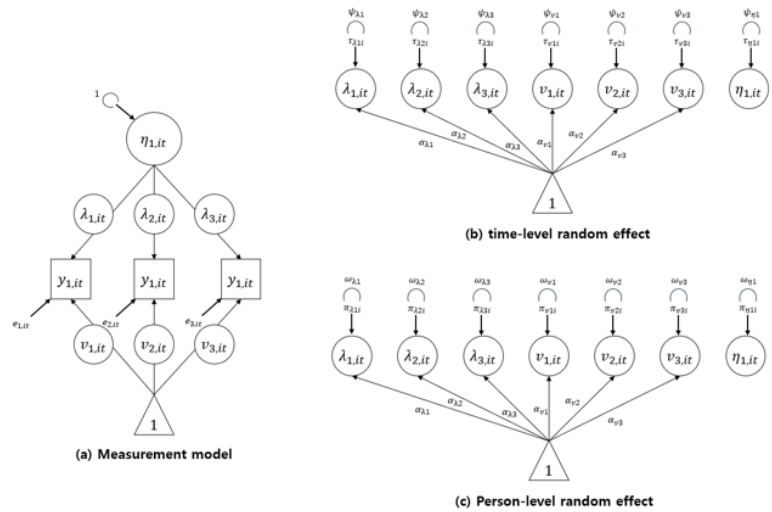


Figure 1. Cross-classified measurement model used for data generation (modified from McNeish et al., 2021).

2. 2 *Simulation Conditions*

In each replication, we generated intensive longitudinal responses from a four-indicator single-factor model for $N = 500$ individuals measured across $T = 40$ occasions. Two groups were used to implement non-homogeneity across conditions, reliability was fixed at .70 (via residual-variance calibration), and the latent factor followed an AR(1) process with $\phi = 0.80$ (McNeish et al., 2021).

Random-effect magnitude varied the cross-classified random effects on loadings and intercepts at the time-level (τ) and person-level (π) under small versus large variance conditions (Asparouhov & Muthén, 2015; Lüdtke et al., 2008). Non-stationarity was induced by time-varying latent moments, including mean growth of 0.02 per occasion, variance increase 0.05 per occasion), or both (McNeish et al., 2021). Non-homogeneity was defined using between-group differences in fixed measurement parameters in loadings ($\Delta\alpha_\lambda = 0.20$), intercepts ($\Delta\alpha_v = 0.30$), or both (Asparouhov & Muthén, 2015; Ryu, 2014). We selected these magnitudes to represent conservative discrepancy from ergodicity while keeping the measurement model realistic for longitudinal CFA. In our results, increasing the magnitude of random effects produced a large shift in longitudinal MI rejection rate, which indicates that these values meaningfully affect the outcomes of interest.

Table 1. Manipulated simulation factors and operational levels

Factor 1. Random effect magnitude		
	F1C1. Small random effect	F1C2. Large random effect
Time-level intercept random effects: $Var(\tau_{vjt})$	.001	.01
Time-level loading random effects: $Var(\tau_{\lambda jt})$	.001	.01
Person-level intercept random effects: $Var(\pi_{vji})$	.100	.30
Person-level loading random effects: $Var(\pi_{\lambda ji})$	.040	.04
Factor 2. Non-Stationary	Factor 3. Non-Homogeneity	

F2C1. Mean growth	+.02 on each occasion	F3C1. Fixed λ diff	$\Delta\alpha_{\lambda}$ = .20 between Groups
F2C2.Variance increasing	+ .05 on each occasion	F3C2. Fixed τ diff	$\Delta\alpha_{v}$ = .30 between Groups
F2C3. F2C1 + F2C2	Both changes applied	F3C3. F3C1 + F3C2	Both diff applied

In Table 1, τ and π denote time- and person-level random effects on intercepts, v, and loadings, λ, with variance specifications applied equally across the four indicators (Asparouhov & Muthén, 2015; Lüdtke et al., 2008). To clarify the conditions, we grouped the factorial design conditions (i.e., data-generating conditions, not separate analytics models) into four scenarios (1) an ergodic scenario varying only random-effect magnitude under stationarity and homogeneity, and three non-ergodic scenarios combining random effects with (2) non-stationarity, (3) non-homogeneity, or (4) both non-stationarity and non-homogeneity.

2. 3 MI testing procedure

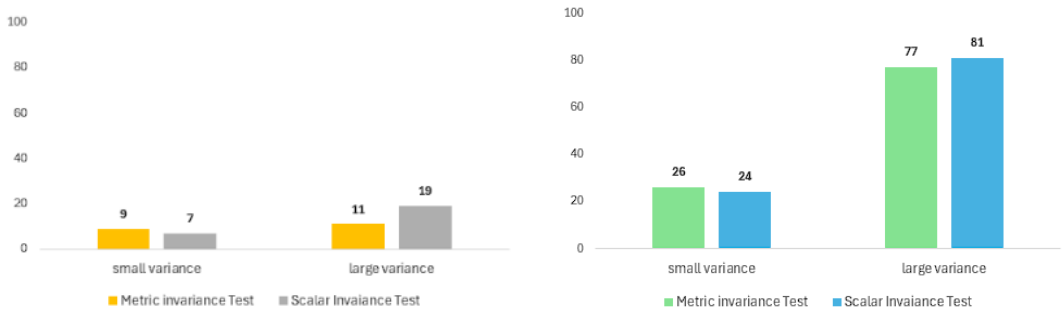
Measurement invariance was tested using a multiple-group CFA, $y = \tau + \Lambda \cdot \eta + \varepsilon$, (Meredith, 1993; Millsap, 2011). We selected t = 1 and t = 30 and fit a two-wave, two-group CFA in which the same four indicators defined a single factor at each wave, allowing the two factors to be correlated.

After assuming configural invariance in this simulation study, we tested metric (equal Λ), and scalar invariance (equal Λ and τ) across groups and waves (Millsap, 2011; Widaman & Reise, 1997). We did not test invariance of residual variances because this constraint is often considered overly restrictive and is not typically required for comparing latent means. Once scalar invariance holds, residual variances may also vary across occasions or groups due to changes in measurement error. Decisions were based on change in fit indices commonly used for approximate invariance (Cheung & Rensvold, 2002; Chen, 2007). For each condition, we report the proportion of replications rejecting metric or scalar invariance (McNeish et al., 2021).

3. Result

3. 1 Ergodic benchmark

Under the random-effects-only benchmark condition, MI rejection rates differed markedly between cross-sectional and longitudinal designs (see Figure 2). With small random-effect variance, the cross-sectional MI test showed low non-invariance, with rejection rates of 9% (metric) and 7% (scalar), whereas the longitudinal MI test exhibited higher sensitivity, with non-invariance of 26% (metric) and 24% (scalar). Increasing the random-effect variance produced little change in the cross-sectional results, which remained low at 11% (metric) and 19% (scalar). On the other hand, the longitudinal MI test responded strongly to the large-variance condition, non-invariance increased sharply to 77% (metric) and 81% (scalar). Overall, these results suggest that longitudinal MI testing is substantially more sensitive to random parameter variability than cross-sectional MI testing, and that larger random-effect magnitudes can induce high rates of apparent non-invariance even in the absence of additional non-stationarity or between-group differences.

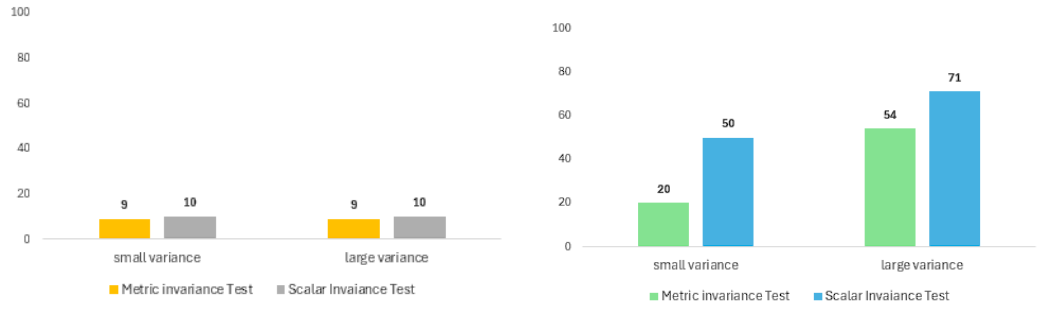


(a) Cross-Sectional MI Testing Result (b) Longitudinal MI Testing Result

Figure 2. Ergodic situation MI Testing reject rate.

3. 2 Non-Ergodic as Non-Stationarity

When non-stationarity was imposed, MI outcomes depended on the type of distributional change and the designs (see Figure 3). Under the combined non-stationarity condition (F2C3), the cross-sectional MI test showed little change relative to the benchmark, with rejection rates remaining at 9% (metric) and 10% (scalar) under both small and large random-effect variance. In contrast, the longitudinal MI test became substantially more sensitive: under small random-effect variance, rejection increased to 20% (metric) and 50% (scalar), and under large random-effect variance it further increased to 54% (metric) and 71% (scalar). These results indicate that longitudinal scalar invariance is particularly affected under mean/variance drift in the present setup, whereas cross-sectional MI remains largely unchanged. For parsimony, Figure 3 reports results for the combined non-stationarity condition (F2C3), rather than presenting the mean-growth-only (F2C1) and variance-increase-only (F2C2) conditions separately.



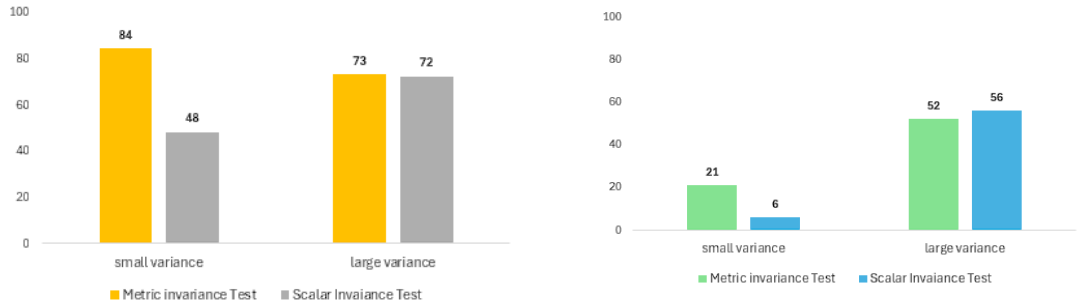
(a) Cross-Sectional MI Testing Result (b) Longitudinal MI Testing Result

Figure 3. Non-Ergodic situation (Non-Stationarity) MI Testing reject rate. (F2C3)

3. 3 Non-Ergodic as Non-Homogeneity

For the non-homogeneity condition, MI results depended primarily on whether between-group shifts were introduced in loadings, intercepts, or both, and the pattern differed across both designs (see Figure 4). Under the combined non-homogeneity condition (F3C3), the cross-sectional MI test produced frequent rejections, with 84% (metric) and 48% (scalar) under small random-effect variance, and 73% (metric) and 72% (scalar) under large random-effect variance. On the other hand, the longitudinal MI test showed comparatively modest rejection under small random-effect variance (21% metric, 6% scalar), but rejection increased under large random-effect variance

to 52% (metric) and 56% (scalar). For simplicity, Figure 4 displays only the combined condition (F3C3), rather than presenting the loading-only (F3C1) and intercept-only (F3C2) conditions separately.

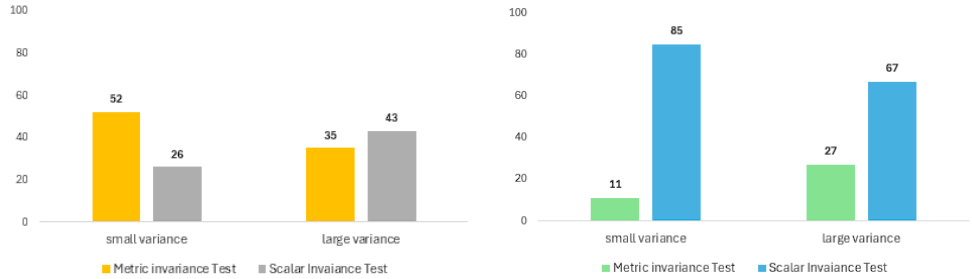


(a) Cross-Sectional MI Testing Result (b) Longitudinal MI Testing Result

Figure 4. Non-Ergodic situation (Non-Homogeneity) MI Testing reject rate. (F3C3)

3. 4 Non-Ergodic as both Non-Stationarity & Non-Homogeneity

When non-stationarity and non-homogeneity were introduced simultaneously, MI rejection rates were lower than those observed when each violation was examined in isolation (see Table A1 in Appendix A and Figure 5). In the cross-sectional design, rejection rates were 52% (metric) and 26% (scalar) under small random-effect variance, and 35% (metric) and 43% (scalar) under large random-effect variance. In the longitudinal design, metric rejection was relatively low at 11% (small variance) and 27% (large variance), whereas scalar rejection was high at 85% (small variance) and 67% (large variance). Notably, relative to the single-violation conditions, the combined condition tended to attenuate metric rejections, especially in the longitudinal design, while scalar rejections remained substantial and, under small random-effect variance, increased sharply. One possible interpretation would be the absorption of the misfits at the configural baseline or become less separable across nested steps when multiple misfits coexist. Or incremental misfit diminishes at some stages while concentrating rejection at others. Because this pattern may depend on the specific parameterization and effect-size settings, it should be interpreted carefully.



(a) Cross-Sectional MI Testing Result (b) Longitudinal MI Testing Result

Figure 5. Non-Ergodic situation (Non-Stationarity & Homogeneity) MI Testing reject rate.

4. Conclusion

This study examined how measurement invariance (MI) testing behaves when longitudinal data are generated under ergodic or non-ergodic conditions. MI outcomes indicated strong design-dependence and the longitudinal design was far more sensitive than cross-sectional design. Large random-effects also yielded high longitudinal rejection rate even under the ergodic benchmark (Asparouhov & Muthén, 2015; McNeish et al., 2021).

Mean drift increased under the scalar non-invariance in the longitudinal design, whereas variance drift had limited impact, suggesting that scalar tests can respond to distributional shifts that mimic intercept change (Meredith, 1993; Millsap, 2011). On the other hand, between-group non-homogeneity in fixed loadings/intercepts showed clear non-invariance in the cross-sectional design, indicating that MI sensitivity depends on whether the design matches the source of non-ergodicity (Fisher et al., 2018; Hamaker, 2012; Molenaar, 2004).

Condition combined with non-stationarity and non-homogeneity did not necessarily increase rejection rates relative to single-violation conditions. This attenuation likely reflects the logic in the nested MI steps. That is, when multiple misfit sources coexist, misfit may be absorbed at the configural baseline or become less separable at later steps, reducing the incremental change used to flag metric/scalar non-invariance. Thus, MI outcomes may be weak diagnostics of the instability source.

Therefore, applied researchers should interpret longitudinal MI failures carefully and consider the models that allow heterogeneity and time variation (e.g., random parameter/multilevel factor models or approximate/Bayesian invariance) (Asparouhov & Muthén, 2015; Lüdtke et al., 2008; Ryu, 2014). Nevertheless, the results of this study do not cover other various situations. As a future study, we plan to conduct simulations that vary effect sizes, sampling designs, and forms of non-stationarity to clarify when MI detects non-ergodic dynamics and when it fails to detect or misclassifies instability. Such a future study would address why the results of this study show that high longitudinal rejection can arise even under an ergodic benchmark when random-effect variance is large.

Reference

Asparouhov, T., & Muthén, B. (2015). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 163–192). Information Age Publishing. https://doi.org/10.1108/978-1-68123-329-120251009

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America, 115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). Guilford Press.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869

McNeish, D., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Molenaar, P. C. M. (1997). Time series analysis and its relationship with longitudinal analysis. *International Journal of Sports Medicine, 18*(Suppl. 3), S232–S237. https://doi.org/10.1055/s-2007-972720

Molenaar, P. C. M. (1999). Longitudinal analysis. In H. J. Ader & G. J. Mellenbergh (Eds.), *Research methodology in the social, behavioural and life sciences* (pp. 143–167). Sage Publications.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 67*(1), 172–194. https://doi.org/10.1111/bmsp.12014

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

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). American Psychological Association. https://doi.org/10.1037/10222-009

Appendix A.

Table A1. Non-Invariance Count Across Ergodic and Non-Ergodic Conditions

	Group				Time			
Test Type	Small		Large		Small		Large	
	Metric	Scalar	Metric	Scalar	Metric	Scalar	Metric	Scalar
Bench Mark	9	7	11	19	26	24	77	81
Non-Stationarity								
F2C1	7	13	9	12	21	58	64	75
F2C2	10	11	10	10	20	12	54	67
F2C3	9	10	9	10	20	50	54	71
Non-Homogeneity								
F3C1	84	17	73	70	20	6	54	57
F3C2	24	28	60	69	16	6	49	55
F3C3	84	48	73	72	21	6	52	56
Non-Stationarity × Non-Homogeneity								
F2C1 × F3C1	76	19	55	54	13	92	39	81
F2C1 × F3C2	16	23	40	60	14	58	45	64
F2C1 × F3C3	27	67	35	43	12	87	74	32
F2C2 × F3C1	59	15	33	48	13	13	32	44
F2C2 × F3C2	12	20	32	49	11	14	28	47
F2C2 × F3C3	52	26	42	43	11	8	35	55
F2C3 × F3C1	52	8	42	41	12	8	33	74
F2C3 × F3C2	12	20	21	49	11	53	37	52
F2C3 × F3C3	52	26	35	43	11	85	27	67