“Physical Self-Concept Changes in a Selective Sport High School: A Longitudinal Cohort-Sequence Analysis of the Big-Fish-Little-Pond Effect” by Herbert W. Marsh, Alexandre J.S. Morin, and Philip D. Parker. Herbert W. Marsh, Alexandre J.S. Morin, and Philip D. Parker are with the Institute for Positive Psychology and Education, Australian Catholic University, Strathfield, NSW, Australia. Address author correspondence to Herbert W. Marsh at herb.marsh@acu.edu.au. http://dx.doi.org/10.1123/jsep.2014-0224

http://dx.doi.org/10.1123/jsep.20__-____

Supplemental Materials for “Physical Self-Concept Changes in a Selective Sport High School: A Longitudinal Cohort-Sequence Analysis of the Big-Fish-Little-Pond Effect”

Contents

1. General Introduction to the Importance of Measurement Invariance

Tests of measurement invariance evaluate the extent to which measurement properties generalize over multiple groups, situations, or occasions. Of particular substantive importance for sport/exercise research are the evaluations of differences across multiple groups (e.g., athlete versus nonathlete; male versus female, age groups, treatment versus control) or over time (i.e., observing the same group of participants on multiple occasions, perhaps before and after an intervention). The need for rigorous tests of whether the underlying factor structure is the same for different groups or occasions has often been ignored in sport/exercise research. However, such comparisons assume the invariance of at least factor loadings and, in some cases, item intercepts (problems associated with differential item functioning). Indeed, unless the underlying factors are measuring the same construct in the same way, and the measurements themselves are operating in the same manner across groups or time, the comparison of parameter estimates is potentially invalid. For example, if gender or longitudinal differences vary substantially for different items used to infer a construct, in a manner unrelated to respondents’ true levels on the latent construct, then the observed differences might be idiosyncratic to the particular items used. From this perspective, it is important to be able to evaluate the full measurement invariance of participants’ responses.

Marsh et al. (2009, 2010, 2014) operationalized a taxonomy of 13 models designed to test measurement invariance that integrates traditional CFA approaches to factor invariance (e.g., Jöreskog & Sörbom, 1988, 1993; Marsh, 1994, 2007; Marsh & Grayson, 1994) with item-response-theory approaches to measurement invariance (e.g., Meredith, 1964, 1993; also see Millsap, 2011, Vandenberg & Lance 2000). Key models test the goodness-of-fit of models with no invariance constraints (configural invariance); invariance of factor loadings (metric or weak invariance), factor loadings and item intercepts (scalar or strong invariance), or factor loadings, item intercepts and item uniquenesses (strict invariance). The final four models (Models 10–13) all constrain mean differences between groups to be zero—in combination with the invariance of other parameters.

Essentially the same logic and taxonomy of models can be used to test the invariance of parameters across multiple occasions for a single group. One distinctive feature of longitudinal analyses is that they should normally include correlated uniquenesses between responses to the same item on different occasions (see Jöreskog 1979, Marsh, 2007, Marsh & Hau, 1996). Although occasions are the most typical test of invariance over a within-person construct like time (i.e., multiple occasions), this is easily extended to include other within-subject variables (e.g., coach or teammates’ ratings of the same athlete); application of the full taxonomy of models is also useful (e.g., Marsh et al., 2009; Marsh et al., 2014; Meredith, 1964, 1993; Millsap, 2011). In the present investigation we focus primarily on the following three models that are central for latent variable models and the evaluation of latent means.

- Configural invariance (whether the a priori factor structure fits when no invariance constraints are imposed over time or groups);
- Metric or weak factorial invariance (tests of the invariance of factor loadings over time and/or groups); and
- Scalar or strong (tests of the invariance of factor loadings and item intercepts over time and/or groups).

We note that it is also possible to test the invariance of the uniqueness terms (including random measurement error) associated with individual items. However, if the focus is on the evaluation of latent relations or latent means based on latent variable models, then uniqueness invariance is not a necessary condition. This follows in that measurement error is controlled in latent-variable models in which each of the constructs is based on multiple indicators (typically, items). Nevertheless, although the valid comparison of latent means does not depend on the invariance of item uniquenesses, there are limitations associated with the noninvariance of item uniquenesses (e.g., manifest means, item variances, and scale variances, as well as relations across manifest constructs, are not comparable across groups). Thus, if the comparison over groups or occasions is based on manifest variables (or scale scores), then the necessary assumptions are the invariance of item uniquenesses, together with the invariance of the factor loadings and item intercepts. Hence, the comparison of manifest variables is considerably more demanding than comparisons based on latent-variable models controlling for measurement error.

2. Cohort Sequential Designs: Longitudinal Invariance Across Samples and Time

As argued by Marsh (1998; Marsh, Craven & Debus, 1998; Parker, Marsh, Morin, Seaton, & Van Zanden, 2014), a multiwave-multicohort design often provides a stronger basis for evaluating developmental differences than cross-sectional comparisons based on many age cohorts, or longitudinal comparisons based on a single age cohort. While sport/exercise, applied, and particularly developmental psychologists often extol the virtues of true longitudinal designs over cross-sectional designs, ultimately, support for the generality of developmental effects requires convergence of results across multiple approaches. Hence, multicohort sequential designs, as in the present investigation, have the advantage of providing tests for history and cohort effects (i.e., based on overlapping data collection waves collected from multiple cohorts of participants) that would not be possible with longitudinal designs based on a single cohort or in cross-sectional designs based on multiple cohorts.

In many longitudinal studies, data are collected from no more than two or three years, making it difficult to fully explore the growth of key educational constructs over the course of major developmental periods such as high school or adolescence. Here we address this issue through the use of cohort sequential designs in which multiple waves of data are collected simultaneously from multiple age cohorts. This strategy provides sport and exercise researchers with a feasible, cost-effective means to explore growth over the course of an entire developmental period (see Brodbeck, Bachmann, Croudace, & Brown, 2013; Enders, 2010; Graham, 2012; Marsh, 1998; Marsh, Craven & Debus, 1998). Furthermore, simulation studies have shown that cohort sequential designs have greater power than standard longitudinal designs when the same number of time waves is collected in each cohort (Graham, 2012). The increasing prevalence of such designs and the development of statistical procedures for analyzing such data are a substantive-methodological synergy (Marsh & Hau, 2007) in which complex substantive issues stimulate the development of stronger methodological tools.

In our demonstration, four waves of data were collected, six months apart, for each of four age cohorts. The cohort-sequential design thus provided a total of 10 waves of data covering five years of high school, in which the multiple waves of data overlap for each successive cohort (Figure 1, repeated here from the in main article shows the structure). One of the most critical aspects of such a design, however, is how to deal with the inevitably large amount of data that is missing by design (white cells in Figure 1), even in the absence of sample attrition and data holes. The advantage of cohort-sequential designs, by contrast, is that the missing time points in all cohorts are missing due to the design of the study, not as a function of participants' characteristics. Thus, these missing data fully correspond to missing-completely-at-random assumptions of modern missing data techniques (Enders, 2010). This suggests that modern missing data techniques can provide unbiased parameter estimates even in the presence of missing data (Enders, 2010). Essentially there are two approaches to estimating growth models with cohort sequential data that attempt to overcome this missing-by-design component. A common approach is to use full information maximum likelihood estimation on data that is stacked and merged across cohorts. In other words, this approach involves reorganizing the data set so that each participant (each line) is specified as having 10 measurement points, with 6 of those being missing (see Figure 1). In this approach, however, some cells of the variance–covariance matrix have zero coverage (i.e., all cases are missing for the entire cohort) and thus, full-information maximum likelihood estimation of these covariances becomes problematic (see Enders, 2010).

Year GroupgPhys		Year 7 Cohort 1	Year 8 Cohort 2	Year 9 Cohort 3	Year 10 Cohort 4
Year 7	M1	M1			
	M2	M2			
Year 8	M3	M3	M3		
	M4	M4	M4		
Year 9	M5		M5	M5	
	M6		M6	M6	
Year 10	M7			M7	M7
	M8			M8	M8
Year 11	M9				M9
	M10				M10

Figure 1 — ~~[AUQ1]~~Cohort sequential design with four cohorts (year in school groups first tested in Years 7–10) and four waves of data for each cohort. Gray squares = Collected data. M1–M10 are 10 latent means that span the 5-year period. White square = Missing by design. Means in each box (solid black rectangles) are matching means based on two different cohorts. Estimates of M3–M8 are each based on results from two cohorts, while those for M1–M2 and M9–M10 are based on a single cohort.

3. Invariance Testing Methodology

As is typical in large longitudinal field studies, a substantial portion (23%) of the sample had missing data on at least one of the four occasions, due primarily to absence, or in some cases to the provision of an illegible (or fictitious) name. All models were fitted using the robust maximum likelihood estimator (MLR) available in Mplus 7.2, in conjunction with multiple imputation procedures to handle missing data. The multiple imputation was based on a data file in which each student had four waves of data, so that imputation was used to fill in missing values within the four waves actually completed by each cohort of students, but not in the data from waves not completed by the cohort (i.e., imputation for the Year 7 cohort had imputed data for the four waves of data in Years 7 and 8, but not for Years 9 and 10). The analyses were done on five imputed data sets and the results were combined automatically by Mplus using the Rubin (1987; Schafer, 1997) strategy to obtain unbiased parameter estimates, standard errors, and goodness-of-fit statistics.

Longitudinal models that evaluate development/change of latent means across multiple waves assume strong/scalar invariance in which factor loadings and item intercepts are assumed to be equal across time waves. In a cohort sequential design, these parameters are thus assumed to be invariant across both time waves and cohorts (thus providing a direct test of possible historical/cohort effects). Furthermore, to estimate the growth trajectories from all time waves and cohorts, cohort sequential designs also assume that overlapping latent means (e.g., T3 and T4 of the Year 7 cohort with T1 and T2 of the Year 8 cohort; see boxes in Figure 1) are invariant across cohorts. Thus, the model constrains loadings and item intercepts to be invariant over time (strong/scalar invariance), and latent means for overlapping time points to be invariant across time points—providing an additional test of cohort-specific historical effects. This is an inherent assumption of cohort sequential designs for models involving latent means, but is rarely tested in applied research.

Goodness of Fit

Historically, applied SEM researchers have sought universal "golden rules" to justify objective interpretations of their data, rather than being forced to defend subjective interpretations (Marsh, Hau, & Wen, 2004). Many fit indices have been proposed (e.g., Marsh, Balla, & McDonald, 1988), but there is even less agreement today than in the past as to what constitutes an acceptable fit. Some still treat the indices and recommended cutoffs as golden rules; others argue that fit indices should be discarded altogether; a few argue that we should rely solely on chi-square goodness-of-fit indices; and many (like us) argue that fit indices should be treated as rough guidelines to be interpreted cautiously in combination with other features of the data. Generally, given the known sensitivity of the chi-square test to sample size, to minor deviations from multivariate normality, and to minor misspecifications, applied SEM research focuses on indices that are sample-size independent (Hu & Bentler, 1999; Marsh, Balla, & Hau, 1996; Marsh, Hau & Grayson, 2005), such as the root mean square error of approximation (RMSEA), the Tucker–Lewis index (TLI), and the comparative fit index (CFI). Population values of TLI and CFI vary along a 0-to-1 continuum in which values greater than .90 and .95 typically reflect acceptable and excellent fits to the data, respectively. Values smaller than .08 or .06 for the RMSEA support acceptable and good model fits respectively.

For the comparison of two nested models, the chi-square difference test can be used, but it suffers from even more problems than the chi-square test for single models; this led to the development of other fit indices (see Marsh, Hau, Balla, & Grayson, 1998). Cheung and Rensvold (2002) and Chen (2007) have suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, then there is reasonable support for the more parsimonious model. Chen (2007) suggests that when the RMSEA increases by less than .015, there is support for the more constrained model. For indices that incorporate a penalty for lack of parsimony, such as the RMSEA and the TLI, it is also possible for a more restrictive model to result in a better fit than a less restrictive model. However, we emphasize that these cutoff values only constitute rough guidelines.

4. Preliminary Invariance Results: Basic Cohort-Sequence Model—Four Cohort Groups and Four waves

We begin with the basic cohort-sequence model depicted in Figure 1. The critical features of this design are that there are four cohorts (year in school) and four waves of data for each cohort. In the first two models (Models M1A and M1B in Table SM1 [see Section 7, herein]), we establish that the a priori correlated uniquenesses are required and result in a substantial improvement in fit (Table SM1). For this reason, correlated uniquenesses are included in all subsequent models. Not surprisingly, given that the physical self-concept scale is brief and has good psychometric properties, the goodness of fit for this model is exceptionally good (e.g., TLI = .979; Table SM1). Inspection of the 96 factor loadings (6 items × 4 waves × 4 cohorts) indicates that all factor loadings are substantial, varying between .65 and .87 in a standardized metric. In summary, there is good support for configural invariance.

In Model M2C (Table SM1) we demonstrate good support for the invariance of factor loadings over four cohorts (between-group cross-sectional comparisons) and four waves (within-person longitudinal comparisons). Indeed, even constraining factor loadings to be invariant over both the four cohorts and the four waves resulted in almost no decrement of fit (e.g., TLI = .980, Model M2C). Hence, there is good support for metric invariance.

In Model M3C (Table SM1) we demonstrate good support for the corresponding tests of invariance of intercepts over four cohorts and four waves. Indeed, even constraining item intercepts to be invariant over both the four cohorts and the four waves resulted in almost no decrement of fit (e.g., TLI = .977, Model M2D). Hence, there is good support for metric invariance.

In a supplemental model (M4A) we also tested for the invariance of uniquenesses over cohorts and waves. However, unlike the corresponding tests of the invariance of factor loadings and intercepts, this model resulted in a moderately large decrement in fit (e.g., TLI = .953, a Δ TLI = .024 compared with Model M3C). We note that the model is borderline in terms of traditional guidelines for testing invariance, but also that it might be possible to achieve a more acceptable fit with partial invariance based on ex post facto relaxation of invariance constraints on the uniquenesses (e.g., Byrne, Shavelson, & Muthén, 1989). However, we chose instead to reject this model, because the invariance of uniquenesses is not an assumption underlying the evaluation of latent means (although it is a necessary condition for the appropriate evaluation of manifest means). Indeed, this is an important advantage of the latent-variable models, that they do not require uniquenesses to be invariant over groups and occasions.

Model M3D is specific to the cohort-sequence design and critical to the interpretation of the results. In particular, it constrains the latent means to be the same with each of the six pairs of matching means (i.e., the matching means in boxes in Figure 1). Thus, for example, the two estimates of M3 based on different cohorts (Wave 3, Year 7 cohort and Wave 1, Year 8 cohort) are constrained to be equal. The latent mean estimates based on Model M3D are shown in Table ~~3~~ in the main text (with boxes representing the matching means constrained to be equal).~~[AUQ2]~~

5. Cohort-Sequence Design of MIMIC Models

An apparently unique aspect of these data is that 38% of the adolescents are elite athletes selected through a highly competitive process, whereas the remaining 62% are nonathletes of similar ages attending the same school and taking mostly the same classes. From this perspective it is of substantive interest to compare developmental stability/change in physical self-concept over time as a function of gender for these two groups of students. Integrating the traditional MIMIC model (treating gender, athletic group, and their interaction as three covariates) with the cohort-sequence analyses presented in the last section provides a substantive-methodological synergy in which we apply new and evolving statistical models to evaluate substantively important issues.

In the first MIMIC model (M5A, Table SM1) we simply added the three MIMIC variables to the final cohort-sequence (M3C with invariance of factor loadings and item intercepts, and matching means for overlapping time points), in which the latent factors (physical self-concept) are predicted by the three covariates (gender: boys vs. girls, athlete group: athletes vs. nonathletes; and the gender-by-athlete-group interaction). In this model, no constraints were placed on the effects of the MIMIC variables, which were freely estimated, and the fit was very good (e.g., TLI = .974). In the next model (M5B, Table SM1), the effects of the MIMIC variables were constrained to be the same within each of the six pairs of matching means (i.e., the matching means in boxes in Figure 1). Thus, for example, the two estimates of gender differences for cell M3, based on different cohorts (e.g., Wave 3, Year 7 cohort and Wave 1, Year 8 cohort) were constrained to be equal. The imposition of this set of 18 constraints (3 MIMIC variables across 6 pairs of cells) resulted in nearly no decrement in fit (e.g., TLI = .974). (The Mplus syntax for MIMIC Models is presented at the end of these Supplemental Materials.)

An important feature of the MIMIC model is parsimony, especially when compared with the corresponding multigroup approach. Thus, for example, in the multiple group approach the 2 (athlete group) × 2 (gender) × 4 (cohort) design would require testing over 16 groups. In the 16 multigroup design it would be very complicated to disentangle invariance associated with each of the three facets (group, gender, and cohort) and their combinations. Furthermore, in the present investigation, the sample sizes of some of the groups would be so small as to make the analyses dubious. Nevertheless, there were critical drawbacks to the use of the MIMIC model, in that researchers cannot easily test the invariance of factor loadings. However, in the MIMIC model it is possible to test the invariance of items' intercepts by freely estimating the effects of the MIMIC variables on all items' intercepts and constraining their effects on the latent means to be zero. Comparison of this model (M5D in Table SM1) with the corresponding model, with the effects of the MIMIC variables on the items' intercepts constrained to be zero and their effects on the latent means freely estimated (M5A in Table SM1), provides a test of intercept invariance across the set of MIMIC variables. If all intercept differences can be explained in terms of a much smaller number of latent means, then there is support for intercept invariance. Results support the invariance of the intercepts, in that M5A fits the data very well (e.g., TLI = .974) but, particularly for indices that control for parsimony, freeing the intercepts does not improve the fit at all (TLI = .972).

In summary, we presented results based on a novel multigroup approach to testing invariance in a cohort sequence. We began by demonstrating good support for the invariance of factor loadings and item intercepts across the four cohort groups and four occasions, spanning five school years. Critically, we then demonstrated support for invariance constraints in latent means across these multiple groups for the overlapping time points (the cells in boxes in Figure 1), resulting in a full model covering five years of high school, even though any one participant only completed data across two years. Next, we extended the typical cohort-sequence design by integrating it with a MIMIC model of the effects of three covariates (gender, athlete group, and their interaction). The effects of these covariates and their interaction with time were explored, based on the traditional orthogonal polynomial contrasts traditionally applied in ANOVAs (but here based on latent rather than manifest means), the other based on latent growth curve models, which is the main substantive focus of the present investigation.

6. References for Supplemental Materials

Also see the references in published main article.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. doi:10.1037/0033-2909.105.3.456

Chen, F.F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. doi:10.1080/10705510701301834

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

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

Jöreskog, K.G. (1979). Statistical estimation of structural models in longitudinal investigations. In J.R. Nesselroade & B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Jöreskog, K.G., & Sörbom, D. (1988). *LISREL 7—A guide to the program and applications* (2nd ed.). Chicago, Illinois: SPSS.

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International.

Marsh, H.W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5–34. doi:10.1080/10705519409539960

Marsh, H.W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum & R.C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774–798). Hoboken, NJ: John Wiley & Sons Inc.

Marsh, H.W., Balla, J.R., & Hau, K.T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.

Marsh, H.W., Balla, J., & McDonald, R.P. (1988). Goodness of fit in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391–410. doi:10.1037/0033-2909.103.3.391

Marsh, H.W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling, 1*(2), 116–146. doi:10.1080/10705519409539968

Marsh, H.W., & Hau, K-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390. doi:10.1080/00220973.1996.10806604

Marsh, H.W., Hau, K-T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220. doi:10.1207/s15327906mbr3302_1

Marsh, H.W., Hau, K-T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu–Olivares & J. McArdle (Eds.), *Contemporary Psychometrics. A Festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H.W., Hau, K.T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471–491. PubMed doi:10.1037/a0019227

Marsh, H.W., Morin, A.J.S., Parker, P., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10,* 85–110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476. doi:10.1080/10705510903008220

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika, 29*(2), 187–206 Retrieved from http://link.springer.com. doi:10.1007/BF02289700

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543 Retrieved from http://link.springer.com. doi:10.1007/BF02294825

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69. doi:10.1177/109442810031002

7. Supplemental Materials (SM) Table

Table SM1 Goodness of Fit for Alternative Cohort-Sequence Models: Global Physical

Model	ChiSq	*df*	RMSEA	CFI	TLI	Description
4-Group Comparisons-No Invariance (Configural)						
M1A	984	1210	0.027	0.979	0.976	No invariance, No correlated uniquenesses
M1B	840	1015	0.025	0.984	0.979	M1A with correlated uniquenesses
4-Group Comparisons-Cohort Invariance of Factor Loadings (Metric)						
M2C[a]	915	1096	0.025	0.983	0.98	M1A with FL invar over Waves and Cohorts
4-Group Comparisons-Cohort Invariance Factor Loadings and Intercepts (Scalar)						
M3C	993	1211	0.026	0.98	0.977	M2C with IN invar over Waves and Cohorts
M3D	996	1212	0.026	0.98	0.976	M3C with IN invar over matching latent means
4-Group Comparisons-Cohort Invariance Factor Loadings, Intercepts Uniquenesses						
M4A	1095	1594	0.038	0.953	0.953	M3C with uniq invar over Waves and Cohorts
M4A1	1077	1495	0.035	0.961	0.96	M4a with Cohort uniq free
M4A2	1071	1472	0.034	0.963	0.961	M3C with Waves uniq free
4-Group Mimic Models						
M5A	1236	1508	0.026	0.977	0.974	M3C with MIMIC = free
M5Ba	1254	1530	0.026	0.976	0.974	M3C with MIMIC invar over matching cells
M5C	1284	1750	0.034	0.96	0.956	M3C with MIMIC = 0
M5D	966	1226	0.027	0.98	0.972	M3C MIMIC LMns = 0 MIMIC Intercepts = free

Note. ChiSq = chi-square; *df* = degrees of freedom ratio; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation. CUs = a priori correlated uniquenesses (uniq) based on the negatively worded items. FL= factor loadings. IN = intercepts. LMns = latent means. Invar = invariance

[a]The Mplus syntax for these models is presented in these Supplemental Materials.

8. Mplus Syntax for Model M5b (See Supplemental Materials Table 1)

```
TITLE: Model M5b (see Table 1 and 4 in main text)
 DATA:
   FILE = wft1234ZGPhyIMPlist.DAT;
   TYPE = IMPUTATION;
 VARIABLE:
 NAMES =
      t1GP1 t1GP2 t1GP3 t1GP4 t1GP5 t1GP6
      t2GP1 t2GP2 t2GP3 t2GP4 t2GP5 t2GP6
      t3GP1 t3GP2 t3GP3 t3GP4 t3GP5 t3GP6
      t4GP1 t4GP2 t4GP3 t4GP4 t4GP5 t4GP6
      elite year sex AGE;

 MISSING ARE ALL (-9);

 usevariables are
  t1GP1 t1GP2 t1GP3 t1GP4 t1GP5 t1GP6
  t2GP1 t2GP2 t2GP3 t2GP4 t2GP5 t2GP6
  t3GP1 t3GP2 t3GP3 t3GP4 t3GP5 t3GP6
  t4GP1 t4GP2 t4GP3 t4GP4 t4GP5 t4GP6
  elite SEX SX_EL;

   GROUPING IS YEAR (7=yR7,8=Yr8,9=Yr9,10=Yr10);

 define:
 center ELITE SEX (GRANDMEAN);
SX_EL = ELITE * SEX;
  ANALYSIS: ESTIMATOR = MLR; !ML; TYPE = BASIC;

MODEL:
  !PHYSICAL SELF-CONCEPT LATENT FACTORS
   (INVARIANT OVER TIME AND COHORT)
 GPHT1 by T1GP1-T1GP6 (L1-L6);
 GPHT2 by T2GP1-T2GP6 (L1-L6);
 GPHT3 by T3GP1-T3GP6 (L1-L6);
 GPHT4 by T4GP1-T4GP6 (L1-L6);
  !Latent variances--free;
   GPHT1 ; GPHT2 ; GPHT3 ; GPHT4 ;
   !PHYSICAL SELF-CONCEPT Correlated residuals
   T2GP1-T2GP6 pwith T1GP1-T1GP6 ;
   T3GP1-T3GP6 pwith T1GP1-T1GP6 ;
   T4GP1-T4GP6 pwith T1GP1-T1GP6 ;
   T3GP1-T3GP6 pwith T2GP1-T2GP6 ;
   T4GP1-T4GP6 pwith T2GP1-T2GP6 ;
   T4GP1-T4GP6 pwith T3GP1-T3GP6 ;
  !PHYSICAL SELF-CONCEPT INTERCEPTS
   (INVARIANT OVER TIME AND COHORT)
   [T1GP1-T1GP6] (INT1-INT6);
   [T2GP1-T2GP6] (INT1-INT6);
   [T3GP1-T3GP6] (INT1-INT6);
        [T4GP1-T4GP6] (INT1-INT6);
  !EFFECTS OF MIMIC COVARIATES
   (INVARIANT OVER TIME AND COHORT)
    GPHT1-GPHT4  on   sex   (sx)    ;
    GPHT1-GPHT4  on   elite (EL)    ;
    GPHT1-GPHT4  on   SX_EL (intS_E)    ;

    MODEL YR7:
       GPHT1 by T1GP1-T1GP6 (L1-L6);
     GPHT2 by T2GP1-T2GP6 (L1-L6);
     GPHT3 by T3GP1-T3GP6 (L1-L6);
     GPHT4 by T4GP1-T4GP6 (L1-L6);
    !Latent variances--free;
     GPHT1-GPHT4*;
     !PHYSICAL SELF-CONCEPT Correlated residuals
     T2GP1-T2GP6 pwith T1GP1-T1GP6 ;
     T3GP1-T3GP6 pwith T1GP1-T1GP6 ;
     T4GP1-T4GP6 pwith T1GP1-T1GP6 ;
     T3GP1-T3GP6 pwith T2GP1-T2GP6 ;
     T4GP1-T4GP6 pwith T2GP1-T2GP6 ;
     T4GP1-T4GP6 pwith T3GP1-T3GP6 ;
     !PHYSICAL SELF-CONCEPT INTERCEPTS
     [T1GP1-T1GP6] (INT1-INT6);
     [T2GP1-T2GP6] (INT1-INT6);
     [T3GP1-T3GP6] (INT1-INT6);
     [T4GP1-T4GP6] (INT1-INT6);
!MEANS (DEFINED TO BE M1-M4 FOR YEAR 7; SEE FIGURE 1)
```

```
 [GPHT1-GPHT4*] (LMM1-LMM4);
 !EFFECTS OF MIMIC COVARIATES
   (DEFINED TO BE M1-M4 FOR YEAR 7; SEE FIGURE 1)
   GPHT1  ON Sex  (sx1);
   GPHT1  ON ELite        (el1);
   GPHT1  ON Sx_el        (ints_e1);
   GPHT2  ON Sex  (sx2);
   GPHT2  ON ELite        (el2);
   GPHT2  ON Sx_el        (ints_e2);
   GPHT3  ON Sex  (sx3);
   GPHT3  ON ELite        (el3);
   GPHT3  ON Sx_el        (ints_e3);
   GPHT4  ON Sex  (sx4);
   GPHT4  ON ELite        (el4);
   GPHT4  ON Sx_el        (ints_e4);

 MODEL YR8:
     GPHT1 by T1GP1-T1GP6 (L1-L6);
     GPHT2 by T2GP1-T2GP6 (L1-L6);
     GPHT3 by T3GP1-T3GP6 (L1-L6);
     GPHT4 by T4GP1-T4GP6 (L1-L6);
    !Latent variances--free;
     GPHT1-GPHT4*;
     !PHYSICAL SELF-CONCEPT Correlated residuals
    T2GP1-T2GP6 pwith T1GP1-T1GP6 ;
    T3GP1-T3GP6 pwith T1GP1-T1GP6 ;
    T4GP1-T4GP6 pwith T1GP1-T1GP6 ;
    T3GP1-T3GP6 pwith T2GP1-T2GP6 ;
    T4GP1-T4GP6 pwith T2GP1-T2GP6 ;
    T4GP1-T4GP6 pwith T3GP1-T3GP6 ;
     !PHYSICAL SELF-CONCEPT INTERCEPTS
     [T1GP1-T1GP6] (INT1-INT6);
     [T2GP1-T2GP6] (INT1-INT6);
     [T3GP1-T3GP6] (INT1-INT6);
     [T4GP1-T4GP6] (INT1-INT6);
! MEANS (DEFINED TO BE M3-M6 FOR YEAR 8; SEE FIGURE 1)
 [GPHT1-GPHT4*] (LMM3-LMM6);
 !EFFECTS OF MIMIC COVARIATES
   (DEFINED TO BE M3-M6 FOR YEAR 8; SEE FIGURE 1)
   GPHT1  ON Sex  (sx3);
   GPHT1  ON ELite        (el3);
   GPHT1  ON Sx_el        (ints_e3);
   GPHT2  ON Sex  (sx4);
   GPHT2  ON ELite        (el4);
   GPHT2  ON Sx_el        (ints_e4);
   GPHT3  ON Sex  (sx5);
   GPHT3  ON ELite        (el5);
   GPHT3  ON Sx_el        (ints_e5);
   GPHT4  ON Sex  (sx6);
   GPHT4  ON ELite        (el6);
   GPHT4  ON Sx_el        (ints_e6);
 MODEL YR9:
    GPHT1 by T1GP1-T1GP6 (L1-L6);
     GPHT2 by T2GP1-T2GP6 (L1-L6);
     GPHT3 by T3GP1-T3GP6 (L1-L6);
     GPHT4 by T4GP1-T4GP6 (L1-L6);
    !Latent variances--free;
     GPHT1-GPHT4*;
     !PHYSICAL SELF-CONCEPT Correlated residuals
    T2GP1-T2GP6 pwith T1GP1-T1GP6 ;
    T3GP1-T3GP6 pwith T1GP1-T1GP6 ;
    T4GP1-T4GP6 pwith T1GP1-T1GP6 ;
    T3GP1-T3GP6 pwith T2GP1-T2GP6 ;
    T4GP1-T4GP6 pwith T2GP1-T2GP6 ;
    T4GP1-T4GP6 pwith T3GP1-T3GP6 ;
     !PHYSICAL SELF-CONCEPT INTERCEPTS
     [T1GP1-T1GP6] (INT1-INT6);
     [T2GP1-T2GP6] (INT1-INT6);
     [T3GP1-T3GP6] (INT1-INT6);
     [T4GP1-T4GP6] (INT1-INT6);
!MEANS (DEFINED TO BE M5-M8 FOR YEAR 9; SEE FIGURE 1)
 !EFFECTS OF MIMIC COVARIATES
   (DEFINED TO BE M5-M8 FOR YEAR 9; SEE FIGURE 1)

 [GPHT1-GPHT4*] (LMM5-LMM8);
   GPHT1  ON Sex  (sx5);
   GPHT1  ON ELite        (el5);
   GPHT1  ON Sx_el        (ints_e5);
   GPHT2  ON Sex  (sx6);
   GPHT2  ON ELite        (el6);
```

```
    GPHT2  ON Sx_el            (ints_e6);
    GPHT3  ON Sex  (sx7);
    GPHT3  ON ELite            (el7);
    GPHT3  ON Sx_el            (ints_e7);
    GPHT4  ON Sex  (sx8);
    GPHT4  ON ELite            (el8);
    GPHT4  ON Sx_el            (ints_e8);
  MODEL YR10:
       GPHT1 by T1GP1-T1GP6 (L1-L6);
      GPHT2 by T2GP1-T2GP6 (L1-L6);
      GPHT3 by T3GP1-T3GP6 (L1-L6);
      GPHT4 by T4GP1-T4GP6 (L1-L6);
     !Latent variances--free;
      GPHT1-GPHT4*;
      !PHYSICAL SELF-CONCEPT Correlated residuals
     T2GP1-T2GP6 pwith T1GP1-T1GP6 ;
     T3GP1-T3GP6 pwith T1GP1-T1GP6 ;
     T4GP1-T4GP6 pwith T1GP1-T1GP6 ;
     T3GP1-T3GP6 pwith T2GP1-T2GP6 ;
     T4GP1-T4GP6 pwith T2GP1-T2GP6 ;
     T4GP1-T4GP6 pwith T3GP1-T3GP6 ;
      !PHYSICAL SELF-CONCEPT INTERCEPTS
      [T1GP1-T1GP6] (INT1-INT6);
      [T2GP1-T2GP6] (INT1-INT6);
      [T3GP1-T3GP6] (INT1-INT6);
      [T4GP1-T4GP6] (INT1-INT6);
!MEANS (DEFINED TO BE M7-M10 FOR YEAR 10; SEE FIGURE 1)
  [GPHT1-GPHT4*] (LMM7-LMM10);
   !EFFECTS OF MIMIC COVARIATES
    (DEFINED TO BE M7-M10 FOR YEAR 10; SEE FIGURE 1)
     GPHT1  ON Sex             (sx7);
     GPHT1  ON ELite           (el7);
     GPHT1  ON Sx_el           (ints_e7);
     GPHT2  ON Sex             (sx8);
     GPHT2  ON ELite           (el8);
     GPHT2  ON Sx_el           (ints_e8);
     GPHT3  ON Sex             (sx9);
     GPHT3  ON ELite           (el9);
     GPHT3  ON Sx_el           (ints_e9);
     GPHT4  ON Sex             (sx10);
     GPHT4  ON ELite           (el10);
     GPHT4  ON Sx_el           (ints_e10);

  MODEL CONSTRAINT:
!CONSTRUCTION OF ORTHOGONAL POLYNOMIAL CONTRASTS AVERAGED ACROSS ALL GROUPS
 new (G1-G10,lin, quad, cubic,quart,GROW,GROW_MSQ,female,elite,FExEL);
 G1=LMM1;G2=LMM2;G3=LMM3;G4=LMM4;G5=LMM5;
 G6=LMM6;G7=LMM7;G8=LMM8;G9=LMM9;G10=LMM10;

 GROW=(G1+G2+G3+G4+G5+G6+G7+G8+G9+G10)/10;

 GROW_MSQ=((G1-GROW)**2+(G2-GROW)**2+(G3-GROW)**2+(G4-GROW)**2+
 (G5-GROW)**2+(G6-GROW)**2+(G7-GROW)**2+(G8-GROW)**2+(G9-GROW)**2+
 (G10-GROW)**2)/10;

  lin=((-.49543369*g1)+(-.38533732*g2)+(-.27524094*g3)+
  (-.16514456*g4)+(-.05504819*g5)+
  (.05504819*g6)+(.16514456*g7)+(.27524094*g8)+
  (.38533732*g9)+(.49543369*g10))/1.175;

  quad=((.52223297*g1)+(.17407766*g2)+(-.08703883*g3)+
  (-.26111648*g4)+(-.34815531*g5)+
  (-.34815531*g6)+(-.26111648*g7)+(-.08703883*g8)+
  (.17407766*g9)+(.52223297*g10) )/1.175;

  cubic=((-.4534252*g1)+(.1511417*g2)+
  (.3778543*g3)+(.3346710*g4)+(.1295501*g5)+
  (-.1295501*g6)+(-.3346710*g7)+
  (-.3778543*g8)+(-.1511417*g9)+(.4534252*g10) )/1.175;

  quart=((.33658092*g1)+(-.41137668*g2)+(-.31788198*g3)+
  (.05609682*g4)+(.33658092*g5)+
  (.33658092*g6)+(.05609682*g7)+(-.31788198*g8)+
  (-.41137668*g9)+(.33658092*g10) )/1.175;


! GENDER: MAIN EFFECT AND ORTHOGONAL POLYNOMIAL
            CONTRASTS OF GENDER DIFFERENCES OVER TIME
 new (F1-F10,Flin, Fquad, Fcubic,Fquart,F_MSQ);
   F1=SX1;F2=SX2;F3=SX3;F4=SX4;F5=SX5;
```

```
  F6=SX6;F7=SX7;F8=SX8;F9=SX9;F10=SX10;
  female=(SX1+SX2+SX3+SX4+SX5+SX6+SX7+SX8+SX9+SX10)/10;

!gender:
   new (F1-F10,Flin, Fquad, Fcubic,Fquart,F_MSQ, F_MSQx);
     F1=SX1/1.175;F2=SX2/1.175;F3=SX3/1.175;F4=SX4/1.175;F5=SX5/1.175;
     F6=SX6/1.175;F7=SX7/1.175;F8=SX8/1.175;F9=SX9/1.175;F10=SX10/1.175;
     female=(F1+F2+F3+F4+F5+F6+F7+F8+F9+F10) /10;

 flin=(-.49543369*F1)+(-.38533732*F2)+(-.27524094*F3)+
 (-.16514456*F4)+(-.05504819*F5)+
 (.05504819*F6)+(.16514456*F7)+(.27524094*F8)+
 (.38533732*F9)+(.49543369*F10);

 fquad=(.52223297*F1)+(.17407766*F2)+(-.08703883*F3)+
 (-.26111648*F4)+(-.34815531*F5)+
 (-.34815531*F6)+(-.26111648*F7)+(-.08703883*F8)+
 (.17407766*F9)+(.52223297*F10);

 fcubic=(-.4534252*F1)+(.1511417*F2)+
 (.3778543*F3)+(.3346710*F4)+(.1295501*F5)+
 (-.1295501*F6)+(-.3346710*F7)+
 (-.3778543*F8)+(-.1511417*F9)+(.4534252*F10);

 fquart=(.33658092*F1)+(-.41137668*F2)+(-.31788198*F3)+
 (.05609682*F4)+(.33658092*F5)+
 (.33658092*F6)+(.05609682*F7)+(-.31788198*F8)+
 (-.41137668*F9)+(.33658092*F10);

F_MSQ=((SX1-FEMALE)**2+(SX2-FEMALE)**2+(SX3-FEMALE)**2+(SX4-FEMALE)**2+
 (SX5-FEMALE)**2+(SX6-FEMALE)**2+(SX7-FEMALE)**2+(SX8-FEMALE)**2+(SX9-FEMALE)**2+
 (SX10-FEMALE)**2)/10;

! ATHLETE GROUP: MAIN EFFECT AND ORTHOGONAL POLYNOMIAL
              CONTRASTS OF GENDER DIFFERENCES OVER TIME
   new (E1-E10,Elin, Equad, Ecubic,Equart,EL_MSQ);
  E1=EL1;E2=EL2;E3=EL3;E4=EL4;E5=EL5;
  E6=EL6;E7=EL7;E8=EL8;E9=EL9;E10=EL10;
 elite=(EL1+EL2+EL3+EL4+EL5+EL6+EL7+EL8+EL9+EL10)/10;
 elin=(-.49543369*E1)+(-.38533732*E2)+(-.27524094*E3)+
 (-.16514456*E4)+(-.05504819*E5)+
 (.05504819*E6)+(.16514456*E7)+(.27524094*E8)+
 (.38533732*E9)+(.49543369*E10);

 equad=(.52223297*E1)+(.17407766*E2)+(-.08703883*E3)+
 (-.26111648*E4)+(-.34815531*E5)+
 (-.34815531*E6)+(-.26111648*E7)+(-.08703883*E8)+
 (.17407766*E9)+(.52223297*E10);

 ecubic=(-.4534252*E1)+(.1511417*E2)+
 (.3778543*E3)+(.3346710*E4)+(.1295501*E5)+
 (-.1295501*E6)+(-.3346710*E7)+
 (-.3778543*E8)+(-.1511417*E9)+(.4534252*E10);

 equart=(.33658092*E1)+(-.41137668*E2)+(-.31788198*E3)+
 (.05609682*E4)+(.33658092*E5)+
 (.33658092*E6)+(.05609682*E7)+(-.31788198*E8)+
 (-.41137668*E9)+(.33658092*E10);   EL_MSQ=((EL1-ELITE)**2+(EL2-ELITE)**2+(EL3-ELITE)**2+(EL4-ELITE)**2+
 (EL5-ELITE)**2+(EL6-ELITE)**2+(EL7-ELITE)**2+(EL8-ELITE)**2+(EL9-ELITE)**2+
 (EL10-ELITE)**2)/10;

! ATHLETE-BY-GENDER INTERACTION: TWO-WAY INTERACTION EFFECT
              AND ORTHOGONAL POLYNOMIAL CONTRASTS OF GENDER DIFFERENCES OVER TIME

  new (I1-I10,Ilin, Iquad, Icubic,Iquart,FEXI_MSQ);
  I1=INTS_E1;I2=INTS_E2;I3=INTS_E3;I4=INTS_E4;I5=INTS_E5;
  I6=INTS_E6;I7=INTS_E7;I8=INTS_E8;I9=INTS_E9;I10=INTS_E10;
 FExEL=(I1+I2+I3+I4+I5+I6+I7+I8+I9+I10)/10;
 ilin=(-.49543369*I1)+(-.38533732*I2)+(-.27524094*I3)+
 (-.16514456*I4)+(-.05504819*I5)+
 (.05504819*I6)+(.16514456*I7)+(.27524094*I8)+
 (.38533732*I9)+(.49543369*I10);

 iquad=(.52223297*I1)+(.17407766*I2)+(-.08703883*I3)+
 (-.26111648*I4)+(-.34815531*I5)+
 (-.34815531*I6)+(-.26111648*I7)+(-.08703883*I8)+
 (.17407766*I9)+(.52223297*I10);

 icubic=(-.4534252*I1)+(.1511417*I2)+
 (.3778543*I3)+(.3346710*I4)+(.1295501*I5)+
```

```
(-.1295501*I6)+(-.3346710*I7)+
(-.3778543*I8)+(-.1511417*I9)+(.4534252*I10);

iquart=(.33658092*I1)+(-.41137668*I2)+(-.31788198*I3)+
(.05609682*I4)+(.33658092*I5)+
(.33658092*I6)+(.05609682*I7)+(-.31788198*I8)+
(-.41137668*I9)+(.33658092*I10); FExI_MSQ=((I1-FExEL)**2+(I2-FExEL)**2+(I3-FExEL)**2+(I4-FExEL)**2+
(I5-FExEL)**2+(I6-FExEL)**2+(I7-FExEL)**2+(I8-FExEL)**2+(I9-FExEL)**2+
(I10-FExEL)**2)/10;

!Non-arbitary intercepts for zero-centered intercept
INT6 = 0 - INT1 - INT2 - INT3 - INT4  - INT5 ;
! traditional approach--FL of first item fixed to 1.0;
L1 = 1.0; ! traditional approach--FL of first item fixed to 1.0;
OUTPUT: stdyx MOD(all, 3) tech1 tech4 SVALUES samp;
```

Author Queries

[AUQ1] The figure will look ~~almost~~ exactly as shown, so please ensure that the terms in the caption “gray squares,” “white squares,” and “solid black rectangles” are specific and unambiguous. It needed, adjustments can be made to the figure. (This is the original table, used as a figure.)

[AUQ2] We do not have a Table 3 at this time. Please advise.