Supplemental Material S1.

Data Handling Procedures

Univariate normality for all items was assessed through inspection of skewness and kurtosis criteria (Table S1), histograms and boxplots of the data (not shown). Study variables did not exceed the recommended thresholds (Kline, 2011) for problematic skew (> 3) or kurtosis (> 10) indicating acceptable univariate normal distribution. Multivariate normality was tested across grades in AMOS 23 using Mardia’s normalized estimate of multivariate kurtosis (1970). In the present study critical ratio of Mardia’s kurtosis exceeded the recommended cut-off value of |5| (Bentler, 2006; Park & Schutz, 2005) for G1 (c.r = 14.087) suggesting violation of multivariate normality (Table S1). Robust techniques were used for handling legitimate outliers and dealing with multivariate non-normality (Kwak & Kim, 2017). Although there are different testing procedures proposed in the literature to assess multinormality (see Kim, 2015), Mardia’s test is commonly used because it is simple and informative regarding non-normality of the data (Zhou & Shao, 2014). Multivariate outliers as indicated by the Mahalanobis distance criterion and its respective p1 and p2 values were not deleted to improve departure from multinormality as this could entail the risk of mistakenly eliminating valid information (Sheskin, 2004, p. 403).

Multivariate collinearity was also examined through the correlation matrix for independent variables (not shown) and the variance inflation factor (VIF). Examination of VIF generated from multiple regression analysis in SPSS 25 by assigning a dummy variable (gender was used) as the dependent variable and task as the independent variable. Given that all correlation coefficients were below < .90 (Hair et al., 2010) and VIF < 10 (Kline, 2011) multicollinearity problems were not addressed further (see Table S2).

Model Testing Procedures and Criteria

The Satorra-Bentler scaled chi-square fit index (also known as the Yuan–Bentler T2 statistic; Yuan & Bentler, 2000) was used to assess overall model fit with a nonsignificant χ^2 value indicating a small discrepancy between expected and observed covariance matrices and thus an acceptable measurement model (Barrett, 2007). Given the sensitivity of χ^2 to sample size (Kline, 1998; Schumacker & Lomax, 2004) we also examined a combination of other fit indices including the comparative fit index (CFI), the non-normed fit index (NNFI, also known as the Tucker-Lewis index or TLI), the root-mean-square error of approximation (RMSEA), the standardized root-mean-square residual (SRMR) and the Akaike information criterion (AIC). CFI analyzes the extent to which the tested model is superior to an alternative model in reproducing the observed covariance matrix (Bentler, 1990). A cut-off value of CFI ≥ .95 is presently recognized as indicative of good fit (Brown, 2006; Hu & Bentler, 1999). TLI evaluates the discrepancy between the chi-square value of the hypothesized model and the chi-square of the null model (Bentler & Bonett, 1980) with values ≥ .95 indicating acceptable fit (Hu & Bentler, 1999). RMSEA assesses how well optimally chosen parameter estimates fit the population covariance matrix (Hooper et al., 2008). RMSEA values ≤ .08 or, more conservatively ≤ .05, suggest acceptable model fit (Brown, 2006). 90% confidence intervals for the RMSEA and results of the closeness of fit test which examines the null hypothesis that RMSEA equals .05 (Browne & Cudeck, 1992) were also reported. SRMR is an index of the average of standardized residuals between the observed and the hypothesized covariance matrices (Bentler, 1995, cited in Chen, 2007). SRMR values as high as .08 are deemed acceptable (Hu & Bentler, 1999) and less than ≤ .05 indicate a well-fitting model (Byrne, 2012; Diamantopoulos & Siguaw, 2000). AIC, a log likelihood measure of fit useful for comparing typically non-nested models, adjusts the chi-square value for the number of estimated parameters (Burnham & Anderson, 2004). AIC's smaller value indicates better fit (Kline, 2013). CFI incremental change was also used in comparison of nested models, with a value < .01 favoring the more parsimonious one (Moran et al., 2013).

In addition to the above model fit indices, the scaled χ^2 difference test ($\Delta\chi^2$) was estimated based on the procedure described by Satorra and Bentler (2001) to compare pairs of nested models. If the $\Delta\chi^2$ test is not statistically significant, the more restricted model is retained as having model fit no worse than the more complex model. If the difference was statistically significant we accepted the more complex model as more adequate. Descriptive fit indices and results of statistical comparisons were not the only criteria we took into account in order to decide the measurement model with the best fit.

References

Barrett, P. (2007). Structural equation modelling: Adjusting model fit. *Personality and Individual Differences, 42*(5), 815–824.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (2006). *EQS 6: Structural equations program manual*. Multivariate Software Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* Guilford.

Burnham, K., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504.

Diamantopoulos, A., & Siguaw, A. D. (2000). *Introducing LISREL: A guide for the uninitiated*. Sage.

Hair, J., Blanck, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Education International.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

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

Kim, N. (2015). Tests based on skewness and kurtosis for multivariate normality. *Communications for Statistical Applications and Methods, 22,* 361–375.

Kline, R. B. (1998). *Methodology in the social sciences: Principles and practice of structural equation modeling*. Guildford.

Kline, R. B. (2011). *Principles and practice of structural equation modelling.* Guildford.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology, 70*(4), 407–411.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrica, 57*(3), 519–530. https://doi.org/10.2307/2334770
Moran, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 1–10). IAP.
Park, I., & Schutz, R. W. (2005). An introduction to latent growth models: Analysis of repeated measures physical performance data. *Research Quarterly for Exercise and Sport, 76*(2), 176–192.
Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrica, 66,* 507–514.
Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Chapman & Hall/CRC.
Schumacker, R. E., & Lomax, R. G. (2004). A beginner's guide to structural equation modeling. Erlbaum.
Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with non-normal missing data. *Sociological Methodology, 30,* 165–200.
Zhou, M., & Shao, Y. (2014). A powerful test for multivariate normality. *Journal of Applied Statistics, 41*(2), 351–363. https://doi.org/10.1080/02664763.2013.839637

Table S1

Variance invariance flator (VIF) across grades

VARIABLE	Prekindergarten	Kindergarten VIF	First grade
NAR	1.59	1.28	1.21
RETEL	1.48	1.44	1.32
ST_COMP	2.04	1.61	1.27
VOC_DEF	2.68	2.08	1.75
CONT_INTER	1.84	1.62	1.83
CONT_RESP	2.85	2.58	1.75
COM_INT	3.16	2.49	1.94
MORPH_DER	1.43	1.37	1.36
MORPH_INFLV	3.61	2.82	2.61
MORPH_INFLN	3.74	2.76	2.18
LIST_COMP	2.20	1.82	1.52
REC_VOC	2.80	1.95	1.64
NAM	2.44	1.82	1.34
INIT_PHON	1.95	1.69	1.51
BL_PHON	2.05	1.83	1.72
PHON_EL	2.39	1.84	1.80

Note. NAR = narrative; RETEL= retelling; ST_COMP = story comprehension; VOC_DEF = vocabulary definition; CONT_INTER = context interpretation; CONT_RESP = context proper response; COM_INT = communicative intent; MORPH_DER = implicit understanding; MORPH_INFLV = implicit understanding of verb inflections; MORPH_INFLN = implicit understanding of noun inflections; LIST_COMP = listening comprehension (sentences); REC_VOC = receptive vocabulary; NAM = naming; INIT_PHON = initial phonemes; BL_PHON = blending phonemes; PHON_EL= phoneme elision.

Table S2.
CFA results of the one-, two- and three-dimensional models by grade

Model	Y-B χ^2	*df*	MLR scaling factor	RMSEA, *p* close (90% CI)	CFI	TLI	SRMR	AIC
Prekindergarten (*n* = 180)								
Model (A)	438.963*	104	1.00	.134, *p* <.001 [.121, .147]	.772	.737	.080	1,3014.55
Model (B)	337.622*	103	.99	.112, *p* <.001 [.099, .126]	.841	.814	.072	1,2910.59
Model (C)	333.299*	101	.99	.113, *p* <.001 [.100, .127]	.842	.812	.066	1,2909.43
Kindergarten (*n* = 269)								
Model (A)	691.196*	104	1.00	.145, *p* <.001 [.135, .155]	.635	.579	.098	1,9799.11
Model (B)	640.734*	103	.99	.139, *p* <.001 [.129, .150]	.666	.610	.102	1,9740.50
Model (C)	648.012*	101	.98	.142, *p* <.001 [.132, .152]	.660	.596	.097	1,9745.28
First-grade (*n* = 351)								
Model (A)	735.327*	104	1.03	.132, *p* <.001 [.123, .141]	.572	.506	.095	2,3930.73
Model (B)	651.283*	103	1.03	.123, *p* <.001 [.114, .132]	.629	.567	.096	2,3842.41
Model (C)	692.753*	100	1.01	.129, *p* <.001 [.120, .138]	.599	.524	.091	2,3869.67

Note. Model (A) = one-dimensional; Model (B) = two-dimensional; Model (C) = three-dimensional; MLR = robust maximum likelihood; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker Lewis index; SRMR = standardized root mean square residual; AIC = Akaike's information criteria.

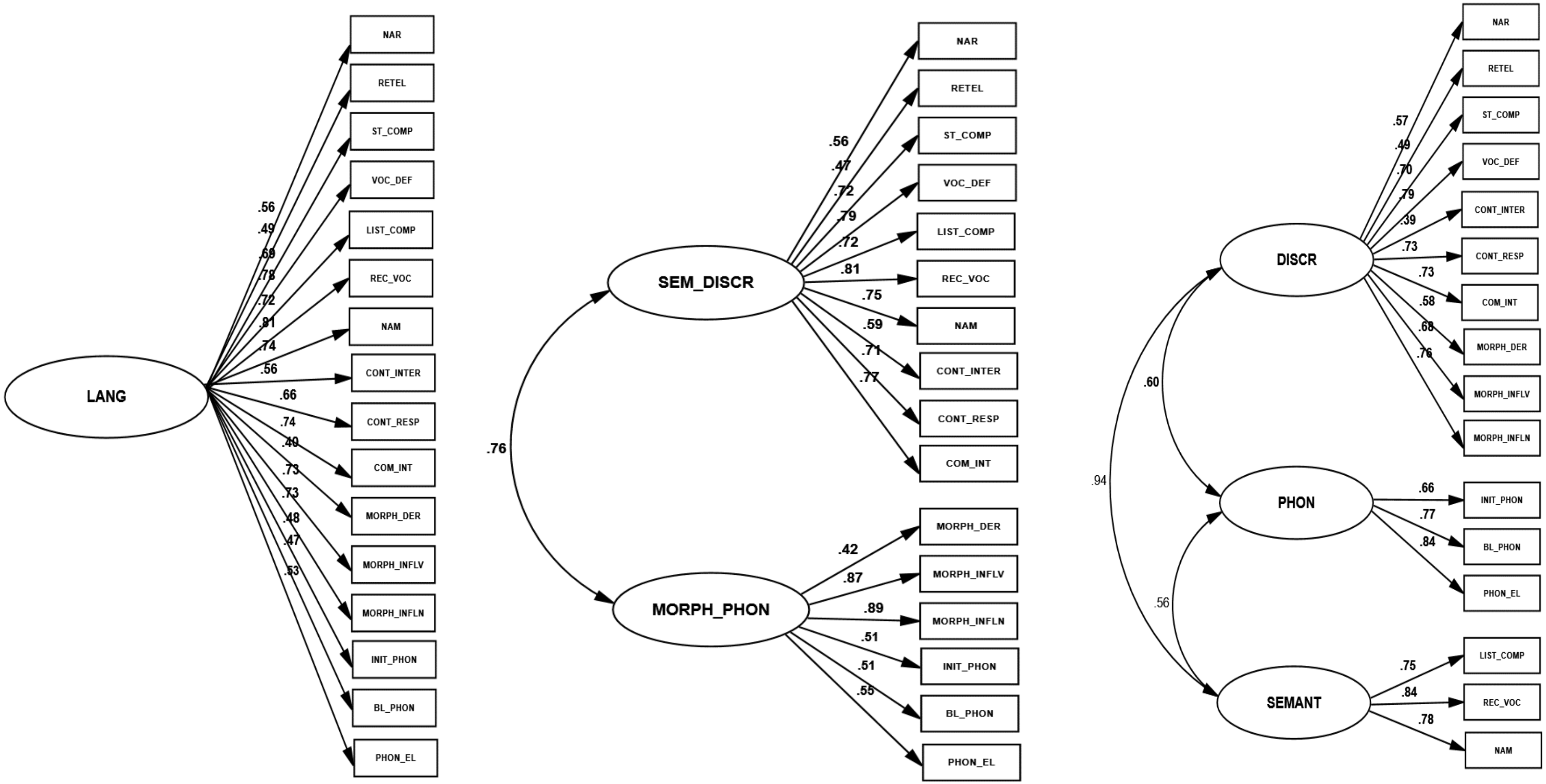

Figure S1. One, two- and three- dimensional models for PK children (standardized factor loadings and factor correlations); LANG = language; SEM_DISCR = semantic_discourse; MORPH_PHON = morphological-phonological; DISCR = discourse; PHON = phonological; SEMANT = semantic; NAR = narrative; RETEL= retelling; ST_COMP = story comprehension; VOC_DEF = vocabulary definition; LIST_COMP = listening comprehension (sentences); REC_VOC = receptive vocabulary; NAM = naming; CONT_INTER = context interpretation; CONT_RESP = context proper response; COM_INT = communicative intent; MORPH_DER = implicit understanding; MORPH_INFLV = implicit understanding of verb inflections; MORPH_INFLN = implicit understanding of noun inflections; INIT_PHON = initial phonemes; BL_PHON = blending phonemes; PHON_EL= phoneme elision.

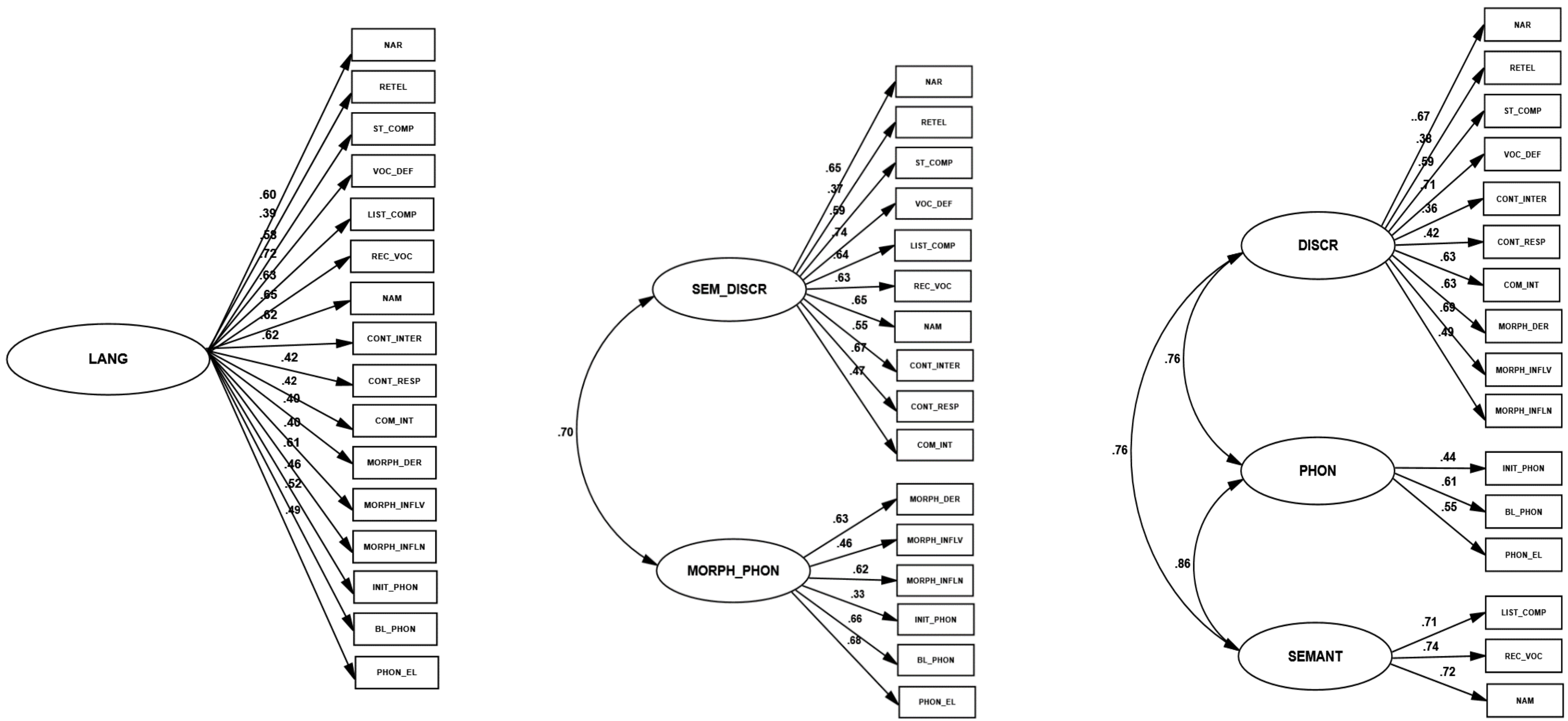

Figure S2. One, two- and three- dimensional models for K children (standardized factor loadings and factor correlations); LANG = language; SEM_DISCR = semantic_discourse; MORPH_PHON = morphological-phonological; DISCR = discourse; PHON = phonological; SEMANT = semantic; NAR = narrative; RETEL= retelling; ST_COMP = story comprehension; VOC_DEF = vocabulary definition; LIST_COMP = listening comprehension (sentences); REC_VOC = receptive vocabulary; NAM = naming; CONT_INTER = context interpretation; CONT_RESP = context proper response; COM_INT = communicative intent; MORPH_DER = implicit understanding; MORPH_INFLV = implicit understanding of verb inflections; MORPH_INFLN = implicit understanding of noun inflections; INIT_PHON = initial phonemes; BL_PHON = blending phonemes; PHON_EL= phoneme elision.

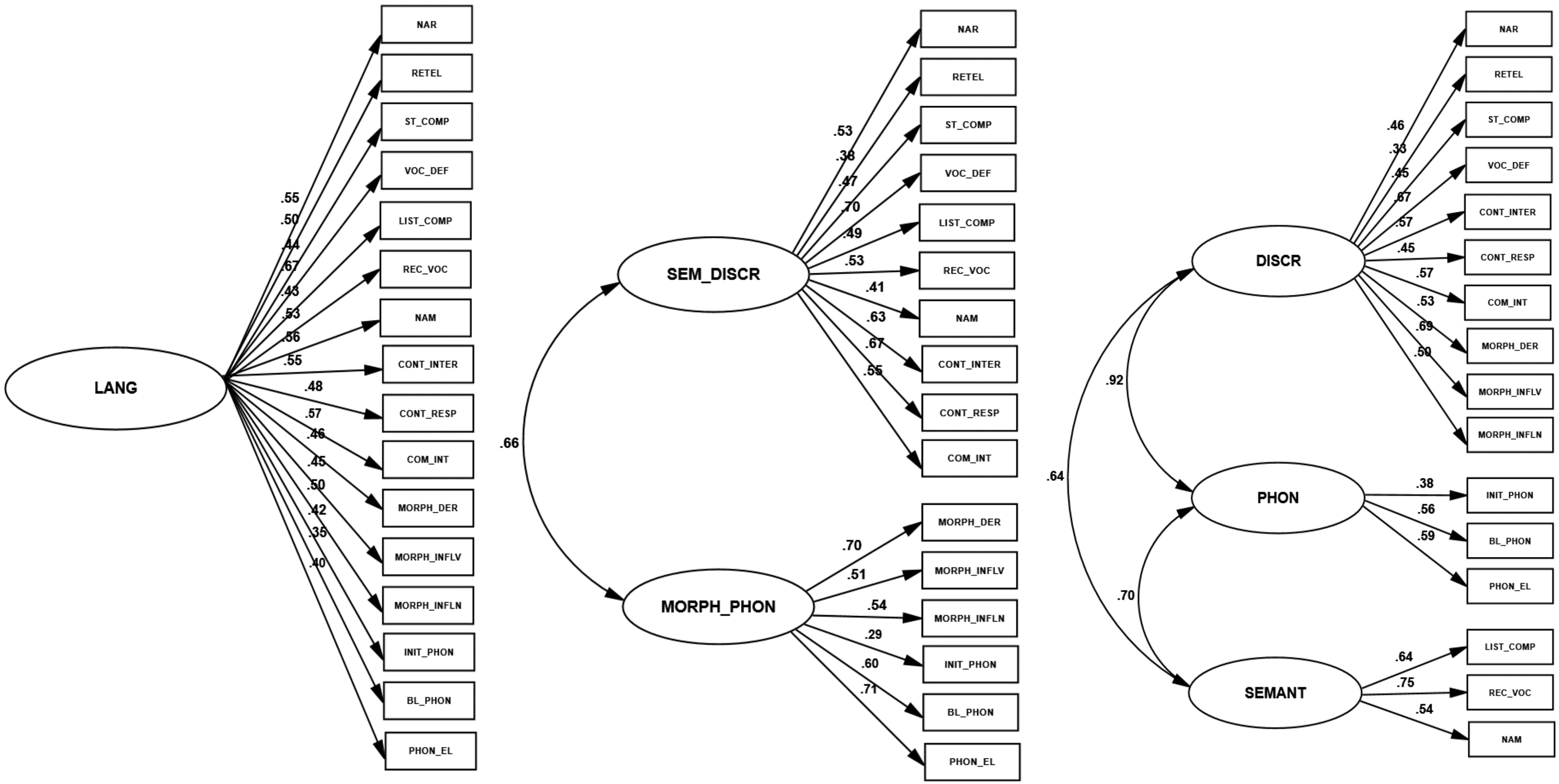

Figure S3. One, two- and three- dimensional models for G1 children (standardized factor loadings and factor correlations); LANG = language; SEM_DISCR = semantic_discourse; MORPH_PHON = morphological-phonological; DISCR = discourse; PHON = phonological; SEMANT = semantic; NAR = narrative; RETEL= retelling; ST_COMP = story comprehension; VOC_DEF = vocabulary definition; LIST_COMP = listening comprehension (sentences); REC_VOC = receptive vocabulary; NAM = naming; CONT_INTER = context interpretation; CONT_RESP = context proper response; COM_INT = communicative intent; MORPH_DER = implicit understanding; MORPH_INFLV = implicit understanding of verb inflections; MORPH_INFLN = implicit understanding of noun inflections; INIT_PHON = initial phonemes; BL_PHON = blending phonemes; PHON_EL= phoneme elision.