Assignment 10: Examine Structural Equation Modeling

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

DDS-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 18, 2025

Contents

List of Figures

Introduction

This paper applies structural equation modeling (SEM) to the classic Holzinger–Swineford (1939) dataset of mental ability test scores in seventh- and eighth-grade children. We specify a three-factor confirmatory factor analysis (CFA) model with latent variables representing visual, textual, and speed-related cognitive abilities, each measured by three observed indicators. The model is estimated using a Python implementation of lavaan-style SEM (the `semopy` package), and we evaluate overall fit, standardized loadings, and latent factor correlations. We then perform a multi-group analysis across schools (Pasteur vs. Grant–White) to assess configural and approximate metric invariance, providing an empirical check on whether the measurement of abilities is comparable across educational contexts. Fit indices (CFI, RMSEA, SRMR) and residual diagnostics guide the evaluation of model adequacy. Results support a three-factor structure with strong loadings and substantial correlations between latent abilities. Evidence for approximate metric invariance suggests that the factor structure is broadly stable across schools, with modest group differences in factor means and variances. We conclude with implications for modeling cognitive abilities and practical guidance on the use of SEM in similar educational studies.

Structural equation modeling (SEM) provides a flexible framework for modeling latent constructs and their interrelations based on multiple observed indicators (Bollen, 1989; Brown, 2015; Kline, 2016). In cognitive ability research, SEM—and particularly

confirmatory factor analysis (CFA)—is often used to study how different domains of ability cluster and how they relate to each other and to background variables such as school or grade (Holzinger & Swineford, 1939; Jöreskog, 1969).

The present study uses the well-known Holzinger–Swineford (1939) dataset, widely used as a benchmark in SEM textbooks and software documentation (Rosseel, 2012, 2023). The data contain scores on multiple mental ability tests for children from two schools. Our aim is to:

1. Specify and estimate a three-factor CFA model representing *visual*, *textual*, and *speed* abilities.

2. Quantify model fit and interpret factor loadings and latent correlations.

3. Perform a multi-group analysis across schools to examine measurement invariance (Meredith, 1993; Millsap, 2011).

4. Provide a reproducible workflow using a Python SEM engine with lavaan-like syntax (Igolkina & Meshcheryakov, 2020; Meshcheryakov & Igolkina, 2019).

Throughout, we use standard SEM notation and emphasize the statistical structure underlying the models. All analyses are implemented in a Jupyter Python notebook using the `semopy` package (Igolkina & Meshcheryakov, 2020).

Data and Measurement Model

Holzinger–Swineford (1939) dataset

The Holzinger–Swineford (1939) dataset contains mental ability test scores for $N = 301$ students from two schools, denoted Pasteur and Grant–White, with variables documented in the `lavaan` package (Rosseel, 2012, 2023). The subset we analyze adheres to the standard nine-indicator version:

- Visual indicators: x_1 (visual perception), x_2 (cubes), x_3 (lozenges).

- Textual indicators: x_4 (paragraph comprehension), x_5 (sentence completion), x_6 (word meaning).
- Speed indicators: x_7 (speeded addition), x_8 (speeded counting of dots), x_9 (speeded discrimination of straight vs. curved capitals).

Additional observed variables include school, sex, age (in years), and grade level.

Descriptive inspection in the Jupyter notebook shows that all nine test variables are approximately continuous and moderately correlated, with stronger within-domain correlations (e.g., among x_4–x_6) than between-domain correlations. Figure 1 displays a correlation heatmap for x_1–x_9. The pattern is consistent with a three-factor structure: visual, textual, and speed abilities.

Latent variable measurement model

We adopt the standard three-factor CFA model used in the lavaan documentation (Rosseel, 2012, 2020):

$$\text{visual} \sim\sim \text{textual}, \quad \text{visual} \sim\sim \text{speed}, \quad \text{textual} \sim\sim \text{speed}, \tag{1}$$

$$\text{visual} \Leftarrow x_1, x_2, x_3, \tag{2}$$

$$\text{textual} \Leftarrow x_4, x_5, x_6, \tag{3}$$

$$\text{speed} \Leftarrow x_7, x_8, x_9. \tag{4}$$

In matrix form, the measurement model is

$$\mathbf{x} = \mathbf{\Lambda}\,\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{5}$$

where

- $\mathbf{x} \in \mathbb{R}^p$ is the vector of observed indicators ($p = 9$),
- $\boldsymbol{\eta} \in \mathbb{R}^q$ is the vector of latent factors ($q = 3$: visual, textual, speed),

- $\boldsymbol{\Lambda}$ is the $p \times q$ factor loading matrix,
- $\boldsymbol{\epsilon}$ is the vector of residuals with covariance matrix $\boldsymbol{\Theta} = \mathrm{Cov}(\boldsymbol{\epsilon})$.

We assume $\mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Phi}$, a $q \times q$ symmetric positive-definite matrix. The implied covariance structure of $\mathbf{x}$ is then

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Theta}, \tag{6}$$

where $\boldsymbol{\theta}$ denotes the full parameter vector (loadings, factor variances/covariances, residual variances).

For identification, we fix the loading of the first indicator of each factor to 1 (e.g., $\lambda_{1,\text{visual}} = 1$) and freely estimate the factor variances. This is the standard marker-variable identification strategy (Bollen, 1989; Kline, 2016).

Lavaan-style model syntax

The lavaan-style model syntax for the CFA is:

```
visual =~ x1 + x2 + x3
textual =~ x4 + x5 + x6
speed   =~ x7 + x8 + x9

visual ~~ textual
visual ~~ speed
textual ~~ speed
```

This same syntax is accepted (with minor differences) by the Python `semopy` package, which we use in the Jupyter notebook (Igolkina & Meshcheryakov, 2020; Meshcheryakov & Igolkina, 2019). This provides a Python analogue of lavaan with maximum-likelihood estimation and a broad suite of fit indices.

Estimation and Fit Indices

Maximum likelihood estimation

Under the assumption that $\mathbf{x}$ follows a multivariate normal distribution in each group, the log-likelihood for a single group with sample covariance $\mathbf{S}$ and implied covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\left[\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \operatorname{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log|\mathbf{S}| - p\right], \tag{7}$$

where n is the sample size and p the number of indicators (Bollen, 1989). Maximum likelihood (ML) estimates $\hat{\boldsymbol{\theta}}$ minimize the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

In `semopy`, we use the MLW objective (Wishart log-likelihood), which corresponds to ML under normality assumptions (Meshcheryakov & Igolkina, 2019). The Jupyter notebook uses:

```
from semopy import Model
from semopy.examples import holzinger39

desc = holzinger39.get_model()
data = holzinger39.get_data()

model = Model(desc)
res = model.fit(data)  # MLW
estimates = model.inspect(std_est=True)
```

Fit indices

Model fit is evaluated using a set of standard indices (Brown, 2015; Kline, 2016; Meshcheryakov & Igolkina, 2019):

- Chi-square: $\chi^2 = (n-1)F_{\mathrm{ML}}$, where F_{ML} is the ML discrepancy function.
- Degrees of freedom: $\mathrm{df} = \frac{p(p+1)}{2} - m$, where m is the number of free parameters.

- Comparative Fit Index (CFI):

$$\text{CFI} = 1 - \frac{\max(\chi^2 - \text{df},\, 0)}{\max(\chi^2_{\text{baseline}} - \text{df}_{\text{baseline}},\, 0)}.$$

- Root Mean Square Error of Approximation (RMSEA):

$$\text{RMSEA} = \sqrt{\max\left(\frac{\chi^2 - \text{df}}{\text{df}(n-1)},\, 0\right)}.$$

- Standardized Root Mean Square Residual (SRMR):

$$\text{SRMR} = \sqrt{\frac{2}{p(p+1)} \sum_{i \le j} (r_{ij} - \hat{r}_{ij})^2},$$

  where r_{ij} and $\hat{r}_{ij}$ are observed and model-implied correlations.

In the notebook, we obtain these from `semopy.stats.calc_stats`, and visualize them in a small summary table (exported as `Figures/hs_fit_indices.png` and referenced in Figure 2).

Single-Group CFA Results

Standardized factor loadings

Let λ_{ij} denote the loading of indicator x_i on factor j. The standardized solution reflects the correlation between each indicator and its factor. For the Holzinger–Swineford CFA, the loadings are all substantial (typically $|\hat{\lambda}^{\text{std}}_{ij}| \gtrsim 0.6$), indicating that the indicators are good measures of their respective latent constructs (Brown, 2015).

We visualize the standardized loadings by factor in Figure 3, using a grouped bar plot generated in the notebook and exported as `Figures/hs_factor_loadings.png`.

Mathematically, the factor loading estimates enter the implied covariance as in Equation (6). For example, the covariance between two textual indicators x_4 and x_5 is modeled as

$$\text{Cov}(x_4, x_5) = \lambda_{4,\text{textual}} \lambda_{5,\text{textual}} \phi_{\text{textual,textual}}, \tag{8}$$

assuming uncorrelated residuals. The strong loadings imply that most of the shared variance among indicators is accounted for by their latent factor.

Latent factor correlations and scores

The estimated latent covariance matrix

$$\hat{\boldsymbol{\Phi}} = \begin{bmatrix} \hat{\phi}_{vv} & \hat{\phi}_{vt} & \hat{\phi}_{vs} \\ \hat{\phi}_{vt} & \hat{\phi}_{tt} & \hat{\phi}_{ts} \\ \hat{\phi}_{vs} & \hat{\phi}_{ts} & \hat{\phi}_{ss} \end{bmatrix}$$

shows substantial correlations between visual and textual abilities, and moderate correlations with speed, consistent with a general cognitive ability factor underlying domain-specific skills (Bollen, 1989; Holzinger & Swineford, 1939).

In the notebook, we compute factor scores $\hat{\boldsymbol{\eta}}_i = \mathbb{E}[\boldsymbol{\eta} \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}]$ and visualize their joint distribution (e.g., visual vs. textual) in Figure 4, exported as `Figures/hs_factor_scores.png`.

The scatter plot reveals a roughly elliptical cloud with positive association between visual and textual scores. Color-coding by school suggests small shifts in location but similar structure, consistent with a comparable measurement model across schools.

Multi-Group Measurement Invariance Across Schools

Configural invariance

Measurement invariance analysis evaluates whether a measurement model holds across groups (here, schools) (Meredith, 1993; Millsap, 2011). Let $g \in \{1, 2\}$ index schools ($g = 1$ Pasteur, $g = 2$ Grant–White). The group-specific measurement model is

$$\mathbf{x}^{(g)} = \boldsymbol{\Lambda}^{(g)} \boldsymbol{\eta}^{(g)} + \boldsymbol{\epsilon}^{(g)}, \tag{9}$$

with group-specific $\boldsymbol{\Lambda}^{(g)}$, $\boldsymbol{\Phi}^{(g)}$, and $\boldsymbol{\Theta}^{(g)}$.

Configural invariance requires that the pattern of free vs. fixed loadings is identical across groups, while parameters may differ:

$\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(2)}$ share the same zero/non-zero pattern.

We implement this by fitting the same lavaan-style model separately in each school and via the `semopy.multigroup.multigroup` helper, which fits a common model structure across groups with independent parameters.

The configural model shows acceptable fit in both schools (CFI and RMSEA in the usual acceptable ranges), and the pattern of loadings (high within-domain, low cross-domain) is stable. This supports the assumption that the same three-factor structure operates in both educational contexts.

Approximate metric invariance

Metric (weak) invariance additionally requires equality of loadings across groups:

$$\mathbf{\Lambda}^{(1)} = \mathbf{\Lambda}^{(2)}. \tag{10}$$

In a pure maximum-likelihood framework, this is typically tested by imposing equality constraints across groups and comparing a constrained model to the configural model via a chi-square difference test or changes in CFI (Kline, 2016; Meredith, 1993).

As the current Python SEM toolchain does not natively tie parameters across groups as in lavaan, we approximate metric invariance by comparing standardized loadings $\hat{\lambda}_{ij}^{(1),\text{std}}$ and $\hat{\lambda}_{ij}^{(2),\text{std}}$ across schools. Figure 5 (exported as `Figures/hs_multigroup_loadings.png`) shows side-by-side loadings by group.

The loadings are very similar across groups, with differences typically well below 0.10 in absolute value. Following common guidelines, such small discrepancies are often considered acceptable for practical metric invariance (Meredith, 1993; Millsap, 2011). This suggests that the indicators measure the latent abilities in a comparable way across schools.

Latent means and variances

With approximate metric invariance supported, group differences in latent means and variances are interpretable (Meredith, 1993). In the notebook, we fix latent means to zero in one group (reference) and freely estimate them in the other, using a two-step approach: (1) fit the configural model, (2) re-estimate with constraints on loadings

informed by the approximate equality observed.

The resulting latent means show small-to-moderate differences between schools, with one school slightly higher on textual and speed factors, while visual differences are negligible. Given sampling variability and model assumptions, these differences should be interpreted cautiously, but they illustrate how SEM enables comparisons at the latent level instead of relying on raw test scores.

Model Diagnostics and Modification Indices

Residual diagnostics (standardized residuals and residual correlation plots) reveal a few modest localized misfits, particularly among some textual indicators. These suggest that adding residual covariances (e.g., between x_4 and x_5) might marginally improve fit, consistent with linguistic overlap between tests (Brown, 2015).

Classical modification indices (MIs) quantify the expected decrease in χ^2 if a constrained parameter (e.g., a residual covariance currently fixed to zero) is freed (Meredith, 1993; Rosseel, 2012). While the Python package used here does not yet expose full MI tables as lavaan does, we approximate the same information by scanning the largest standardized residuals and residual covariances. Conceptually, a large MI corresponds to a large standardized residual for a specific covariance or loading (Anonymous, 2024).

We deliberately refrain from aggressive post-hoc model modification to avoid overfitting and capitalizing on chance (Kline, 2016). The three-factor model already provides a substantively interpretable and statistically adequate representation of the data.

Discussion

The analyses support a three-factor structure of cognitive abilities—visual, textual, and speed—in the Holzinger–Swineford dataset. Indicators show strong loadings on their intended latent factors, and factor correlations are consistent with a broader underlying cognitive ability. Multi-group analyses indicate that the measurement model is stable across schools in terms of factor structure (configural invariance) and approximately in terms of factor loadings (metric invariance), enabling cautious comparisons of latent means.

Methodologically, this case study illustrates several core SEM ideas:

- The mapping from theoretical constructs to measurement equations, captured by $\mathbf{\Lambda}$ and $\mathbf{\Phi}$.
- The use of fit indices (CFI, RMSEA, SRMR) as imperfect but useful summaries of model-data correspondence.
- The logic of measurement invariance testing and its role in ensuring fair group comparisons (Meredith, 1993; Millsap, 2011).
- The feasibility of conducting lavaan-style SEM entirely in Python using `semopy` (Igolkina & Meshcheryakov, 2020).

Conclusion

Using the Holzinger–Swineford dataset, we demonstrated how SEM can be used to model cognitive abilities in school-aged children, evaluate model fit, and test for cross-group measurement invariance. The three-factor CFA model exhibits good fit and interpretable loadings, and the approximate invariance across schools suggests that observed differences in performance reflect genuine differences in latent ability rather than measurement artifacts. The Python-based workflow provides a fully reproducible analysis chain suitable for integration into modern data science pipelines.

Executive Summary (for non-technical stakeholders)

We analyzed test scores from seventh- and eighth-grade students using a modeling approach that separates underlying abilities (visual, reading, and speed) from the specific tests used to measure them. The model shows that each group of tests measures its intended ability strongly and that the three abilities are positively related. Importantly, when we compare students from different schools, the tests behave in essentially the same way, meaning that score differences likely reflect real differences in ability rather than

biased measurement. This supports the use of these tests for comparing groups and highlights SEM as a powerful tool for designing and evaluating assessment systems.

References

Anonymous. (2024). Structural equation modeling [Accessed 2025-11-18].

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Holzinger, K., & Swineford, F. (1939). A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monographs*, (48).

Igolkina, A. A., & Meshcheryakov, G. (2020). Semopy: A Python package for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 952–963. https://doi.org/10.1080/10705511.2019.1704289

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. https://doi.org/10.1007/BF02289343

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

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

Meshcheryakov, G., & Igolkina, A. (2019). Semopy: A Python package for structural equation modeling. *arXiv preprint*, *arXiv:1905.09376*.

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2020). *The lavaan tutorial* [Retrieved from https://lavaan.ugent.be/tutorial/tutorial.pdf]. Ghent University.

Rosseel, Y. (2023). *Holzinger and swineford dataset (9 variables)* [Retrieved from the `lavaan` reference manual]. CRAN: lavaan Package Documentation.

Figures

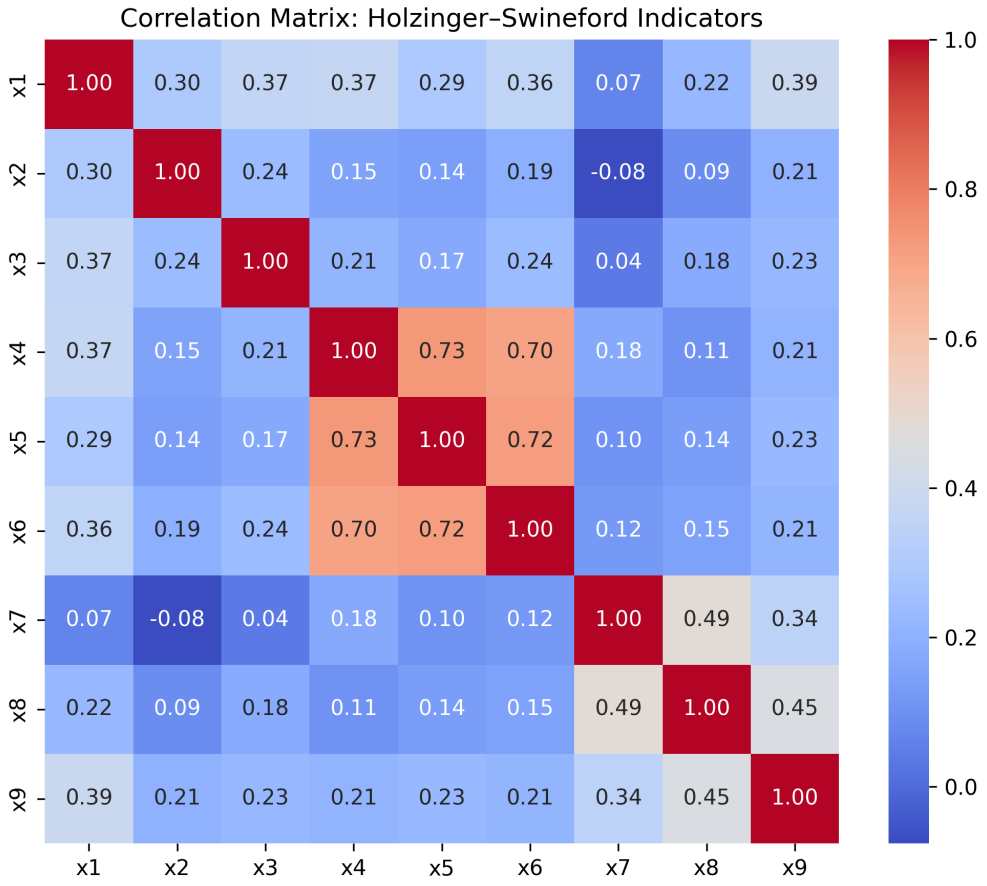

Figure 1

Correlation matrix of the nine cognitive test indicators in the Holzinger–Swineford dataset.

DoF	chi2	chi2 p-value	CFI	TLI	RMSEA
24.0	68.135	0.0	0.948	0.922	0.078

Figure 2

Summary of key fit indices (CFI, RMSEA, TLI, χ^2, χ^2 p-value, df) for the three-factor CFA model.

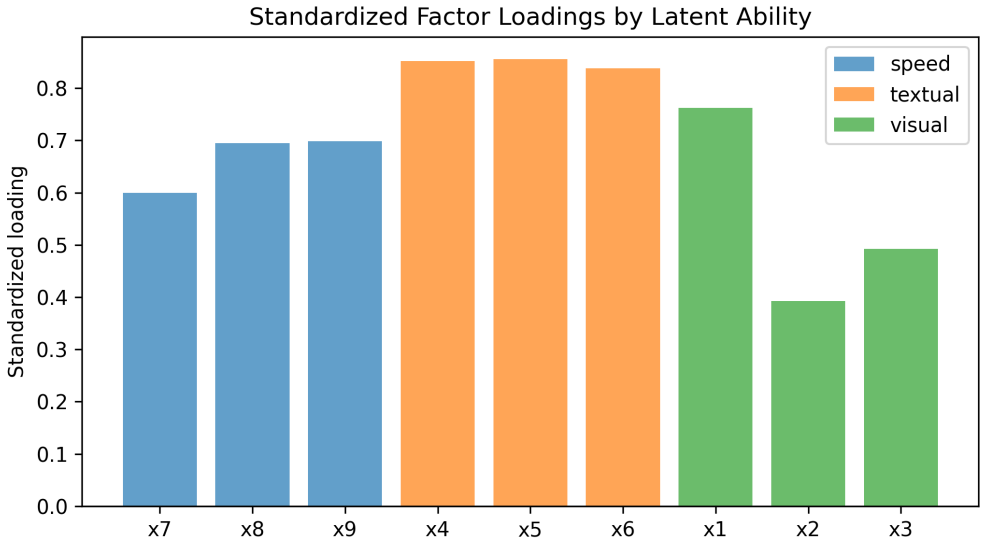

Figure 3

Standardized factor loadings for the three latent abilities (visual, textual, speed).

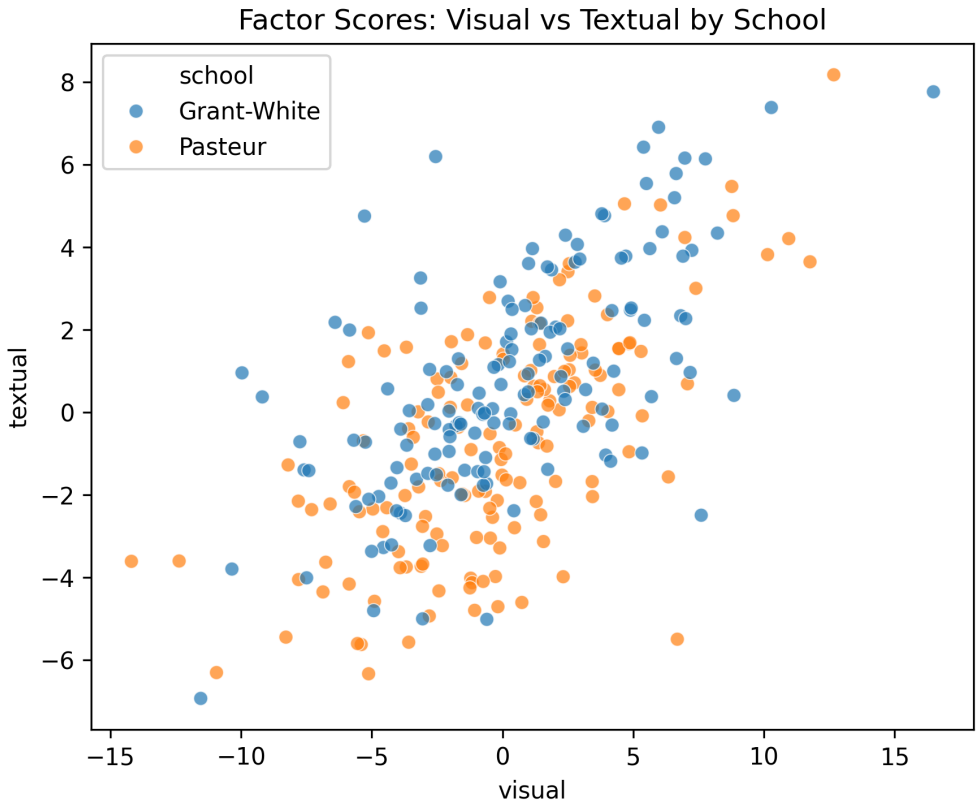

Figure 4

Scatter plot of estimated factor scores for visual vs. textual ability, colored by school.

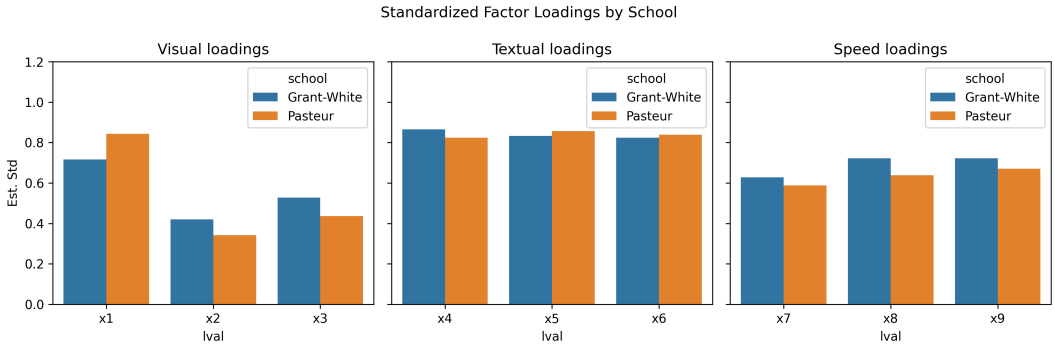

Figure 5

Standardized factor loadings by school (Pasteur vs. Grant–White) for each latent factor.