An Analysis of Massachusetts' Standardized Testing through Multi-Group Structural Equation Modelling

Landon Hurley
Psychology, SUNY Geneseo

Herb Susmann
Mathematics, SUNY Geneseo

March 24, 2014

1   Abstract

The Massachusetts Comprehensive Assessment System (MCAS) is a set of standardized tests administered to Massachusetts public school students. We performed a series of statistical analyses on a subset of 10th grade student test results in English, Mathematics, and Biology tests. An important property of any testing procedure is that of measurement invariance: the test should measure the same underlying constructs, in the same manner, regardless of what groups (ethnicity, gender, etc.) each testee belongs to. We employed Structural Equation Modeling (SEM) to test for measurement invariance across four ethnic groups (White, Black, Hispanic, and Asian) in the MCAS results. Our results indicate that the MCAS does not function similarly across the ethnic groups, a result which undermines the validity of the test. Additionally, we applied multivariate regression and multivariate regression trees to analyze results from a questionnaire given to testees. Our results indicate that students who indicate ambition to pursue a bachelor's degree have higher scores on each test and are significantly affected by higher quality educational programs.

2   Introduction

2.1   Data

The Massachusetts Comprehensive Assessment System (MCAS) is a standardized test administered to Massachusetts public school students in grades 3-10 since 1998. We examine a subset sample of 10th grade students' results from the 2009 Spring MCAS test. The data also contains basic demographic information for each student comprised of race/ethnicity, gender, and an academic engagement questionnaire. Each test consists of multiple choice items and open ended type questions that are graded on a holistic scale, implemented as a 5 point integer scale.

The examinations are comprised of three separate knowledge domains: English, Mathematics and a science component. Every student takes identical versions of the English and Mathematics sections; the science component can be fulfilled by taking either a Biology, Chemistry, Introductory Physics, or Technology/Engineering test. (*Spring 2009 MCAS Tests: Summary of State Results*, 2009)

The MCAS examination structure has received substantial criticism from both educational activists and theoreticians. The former, comprised mainly of teachers and educational policy analysts,

claim that the test is unfairly biased against students as a function of demographics, while citing the significant impediment that poor performance upon the tests has upon both immediate and long-term outcomes. (Gaudet, 2000) Theoretical perspectives on test design are the primary domain of psychometrics, who hold the unfortunate distinction of both supporting standards-based assessment, and subsequently watching their recommendations be ignored when put into practice, either through misinterpretation of statistical results, or the implementation of unrealistic expectations that ignore the original purpose: measurement, or the designation of numerical scales to classify underlying unobservable latent constructs.

2.2 Theoretical structure of the test

Our paper begins with an analysis of the psychometric equivalence of the MCAS test within our sample of 10th grade students, followed by an investigation into performance within each knowledge domain, conditioned upon the demographics and item responses found within a questionnaire included with the test.

3 Methodology

3.1 Subjects

Our sample is comprised of 10,515 10th grade students enrolled in Massachusett's public education system, and is assumed to be drawn as an unbiased sample of the total universe of 10th graders. Provided demographics present as follows: 4.48% Asian, 2.97% Black, 3.63% Hispanic/Latino, 1.8% multiracial, 0.17% Native American, .12% Pacific Islander, and 86.82% White; 50.32% was female, and there are no significant interactions between gender and ethnicity.

Unfortunately for our purposes, the primary point of contention regarding exam disparity, student socio-economic status (SES), was not available, which limits our ability to test and explore the purported causal relationship with student scores (Gaudet, 2000), as mediated by ethnicity or spatial location.

3.2 Test Validity

A critical assumption within standardized testing, and in truth any measurement scale, is that the test demonstrates a psychometric property called measurement invariance/equivalence. This property is a series of mathematical constraints applied to latent variables that are assumed to be non-significantly different between groups. Latent variables are unobservable constructs that underlie the observed measures; one typically estimates them by computing standard principal axis factor analysis. The constraints serve to empirically demonstrate that across sub-populations, the test measures the same concepts, and that each item tests in a similar way the same construct in comparison to the overall sample. This is shown by imposing restrictions in a sequential four phase procedure:

1. Configural invariance: Tests whether the same factor model (i.e., latent variables, equivalent to knowledge subdomains, for example grammar) is found within each subgroup.

2. Weak invariance: Tests whether in addition to the same factor structure, the same items load equivalently onto the same structure.

3. Strong invariance Tests that the intercepts, item loadings, and factor structure are equivalent across groups.

4. Strict invariance: Tests the assumption that in addition to the preceding steps, the residual variances are equivalent across groups.

These four factors are argued to be necessary conditions for a fair and equitable comparison(Meredith, 1993), and are established by way of multi-group confirmatory factor analysis (MG-CFA). In this paper, we examine these factors using a series of R packages (Revelle, 2014; Rosseel, 2012; Pornprasertmanit, Miller, Schoemann, & Rosseel, 2013). Exploratory bootstrapped oblimin factor analyses were conducted upon each test to establish the number of domains and high loading items within each of the three tests. As a consequence of the extreme ethnicity imbalance within three of the four science exams, only Biology was tested using MG-CFA.

3.3 Questionairre Results

We investigated the information found within the student questionnaire by relating them to the students' scores on the three sections. We used both multivariate parametric regressions and a non-parametric predictive procedure using the mvpart package, which models multivariate regression trees.

Missing data in survey and test data is an important consideration, and there exist a number of techniques to produce more robust solutions; most work under the Bayesian estimation process for missing data pioneered by Donald Rubin(Rubin, 2009). However, within the structure of the exam, missing data was unsurprisingly a rare occurrence: the highest proportion of missing for any item was .008, and 91.23% had no missing questions. Even so, for the MG-CFA we attempted to account for this by utilizing the Full Information Maximum Likelihood (FIML) estimation in conjunction with a large number of bootstraps to obtain accurate estimates of asymptotic errors. However, the questionnaire exhibited extreme proportions of non-response, which is problematic for the multivariate regression techniques as it produces potentially biased results under the assumption that items are missing completely at random. As such, we processed the original responses using the mice package(van Buuren & Groothuis-Oudshoorn, 2011) using 13 imputed data frames with 15 iterations in each dataframe. Regression trees are relatively robust to missing data because all information is not processed simultaneously, but rather sequentially in order of variable importance. As such we did not take any additional steps when conducting that analysis.

4 Results

4.1 Exploratory factor analyses

Given the preceding information, we examined the Mathematics, English, and Biology sections of the exam using a oblimin rotated factor structure, with the number of latent factors equal to the number of general domains. Factor loadings were calculated using maximum likelihood estimates from the tetrachoric correlation matrices for each subset of items. The psych package was used to calculate the factor structure. Items that loaded higher than .3 on a factor were specified as indicators, and diagrams of each factor can be located in Appendix A.

4.2   MG-CFA

As a consequence of the unequal numbers of test takers between Biology and the rest of the science exam options, we chose to run two separate tests for measurement invariance: the first was composed of the total sample on the English and Mathematics, and the second was composed of the Biology testees (N=9,239). Diagrams representing the EFA results can be found in Appendix A.

The latent variable models for each test were specified using a FIML procedure with bootstrapped sampling within the four primary ethnicities: A, B, H, W. Other ethnicities were excluded because of their relative rarity within the sample. The χ^2 goodness of fit statistic was used to evaluate the fit of the model. However, this statistic, defined as

$$\chi^2_{d.f} = 2(N-1) \cdot F_0 \tag{1}$$

scales as a function of both sample size and the minimum function test statistic F_0, traditional measurement invariance, which uses changes in significance to reject a constraint, fails here as the sample size guarantees significant results. In response to this fact, we employ three different fitness criterion: the Δ Comparative Fit Index (CFI) (Bentler, 1990), the Δ root mean square error of approximation (RMSEA), and the Δ Tucker-Lewis Index (TLI) (Chen, 2007). For all three, decreases in model fit of greater than .01 are considered significant reductions in model fit. In addition, we include An Akaike Information Criterion (AIC), to demonstrate whether model fit increases or decreases between sequential steps in comparison to the theoretical true model.

4.3   English

The English section was constructed using the system of latent and test items presented below as predictors of the unstandardized English score.

Latent Variable	Latent Variable Indicators
1	8, 10, 11, 12, 13, 16, 17, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 36, 37
2	27, 35, 9, 18
3	38, 39, 40

All latent variable indicators are questions from the English section of the test (e.g. '8' represents Question 8, column 'eitem8' in the provided dataset).

Configural Invariance

```
##    chisq      tli      cfi    rmsea      aic
## 3.981e+03 9.470e-01 9.510e-01 2.400e-02 2.154e+05
```

As the summary statistics given in the analyses show, configural invariance is supported, meaning that all four ethnicities employ the same conceptual frameworks to answer the test items.

Metric (Weak) Invariance

Weak Invariance does impact the model fit compared to Configural Invariance, however negligibly: ΔCFI = .002, ΔRMSEA = 0.000 , $\Delta\chi^2$=187.875, and the ΔTLI = 0.000.

```
##     chisq       tli       cfi     rmsea        aic
## 4.150e+03 9.470e-01 9.490e-01 2.400e-02 2.154e+05
```

Strong Invariance

Strong Invariance is established using the model fit indexes above and beyond weak invariance: ΔCFI = .002, ΔRMSEA = .000 , and the ΔTLI =.000 and $\Delta\chi^2$=187.875.

```
##     chisq       tli       cfi     rmsea        aic
## 4.337e+03 9.470e-01 9.470e-01 2.400e-02 2.154e+05
```

Strict Invariance

Strict invariance does not hold, with a significant ΔCFI = .034, ΔRMSEA = .006, $\Delta\chi^2$=1670.297, and the ΔTLI = .03 indicating lack of fit.

```
##     chisq       tli       cfi     rmsea        aic
## 6.008e+03 9.180e-01 9.120e-01 3.000e-02 2.169e+05
```

4.4   Mathematics

While the EFA presented a good model fit for a five factor model, when the SEM model was imposed upon the data it failed to converge due to underspecification. As such, factor 5 was removed because it consisted of only four low loading items. The latent variable model for Mathematics is presented in the table below.

Latent Variable	Latent Variable Indicators
1	4, 5, 8, 11, 12, 14, 25, 32, 35, 37, 40, 42
2	6, 10, 17, 24, 39
3	22, 28, 34, 38, 41
4	18, 20, 21, 31, 42

All latent variable indicators are questions from the Mathematics section of the test (e.g. '8' represents Question 8, column 'mitem8' in the provided dataset).

Configural Invariance

```
##     chisq       tli       cfi     rmsea        aic
## 2.704e+03 9.640e-01 9.690e-01 2.800e-02 3.160e+05
```

The summary statistics presented below show that configural invariance is supported, meaning that all four ethnicities employ the same conceptual frameworks to answer the test items.

Metric (Weak) Invariance

Weak Invariance does impact the model fit compared to step 1, however negligibly: ΔCFI = .004, ΔRMSEA = .001 , $\Delta\chi^2$=260.076, and the ΔTLI = .002.

```
##     chisq       tli       cfi     rmsea       aic
## 2.965e+03 9.630e-01 9.650e-01 2.900e-02 3.161e+05
```

Strong Invariance

Strong Invariance does not substantively change model fit above and beyond weak invariance: ΔCFI = .002, ΔRMSEA = .000, the ΔTLI = .000 and $\Delta\chi^2$=167.652.

```
##     chisq       tli       cfi     rmsea       aic
## 3.132e+03 9.630e-01 9.630e-01 2.900e-02 3.162e+05
```

Strict Invariance

Strong Invariance imposes a significant constraint upon model fit: ΔCFI = .017, ΔRMSEA = .005, the ΔTLI = .014 and $\Delta\chi^2$=998.348.

```
##     chisq       tli       cfi     rmsea       aic
## 4.131e+03 9.490e-01 9.460e-01 3.300e-02 3.170e+05
```

4.5   Biology

We proceeded with the same methodology to test the science section in isolation, using identical Δ constraints as above. The test for configural invariance produced non-significant results, indicating general acceptance of the factor structure. In addition, each subsequent test produced similar results, indicating that across all four primary ethnicities, the test operated with equivalent measurement. Biology presented an interpretable factor structure, which was represented with the following mathematical structure.

Latent Variable	Latent Variable Indicators
1	1, 2, 13, 14, 15, 22, 27, 30, 33, 34, 37, 40, 41
2	3, 5, 18, 19, 28, 31, 35, 43, 45
3	8, 10, 11, 12, 23, 24, 25
4	4, 23, 29, 32, 44

All latent variable indicators are questions from the Biology section of the test (e.g. '8' represents Question 8, column 'sitem8' in the provided dataset).

Configural Invariance

The analyses' summary statistics show that configural invariance is supported, meaning that all four ethnicities employ the same conceptual frameworks to answer the test items.

```
##     chisq       tli       cfi     rmsea       aic
## 3.981e+03 9.470e-01 9.510e-01 2.400e-02 2.154e+05
```

Metric (Weak) Invariance

While Weak Invariance does impact the model fit compared to step 1, the change is negligible: ΔCFI = .002, ΔRMSEA = .000 , $\Delta\chi^2$=168.330, and the ΔTLI = .000.

```
##     chisq       tli       cfi     rmsea       aic
## 4.150e+03 9.470e-01 9.490e-01 2.400e-02 2.154e+05
```

Strong Invariance

Strong Invariance fails to substantively change model fit above and beyond metric invariance: ΔCFI = .002, ΔRMSEA = .000 , and the ΔTLI = .000 and $\Delta\chi^2$=187.875.

```
##     chisq       tli       cfi     rmsea       aic
## 4.337e+03 9.470e-01 9.470e-01 2.400e-02 2.154e+05
```

Strict Invariance

Strict Invariance does substantively change model fit above and beyond strong invariance: ΔCFI = .034, ΔRMSEA = .006 , and the ΔTLI = .03 and $\Delta\chi^2$=1670.297.

```
##     chisq       tli       cfi     rmsea       aic
## 6.008e+03 9.180e-01 9.120e-01 3.000e-02 2.169e+05
```

4.6   Questionnaire

4.6.1   Multivariate Regression Tree

Without preexisting beliefs about the nature of the relationship between the response surface and the questionnaire and demographics, we chose to begin the analyses with an undirected multivariate tree regression (De'Ath, 2002) which is conceptually similar to running a multinomial regression upon the different clusters of the response variables. The tree also tests by default the possibility that which science test a student takes interacts with other outcomes –in addition, forcibly partitioning the data does not meaningfully increase model fit. We also made the decision to leave non-response as a legitimate factor level: this allows us to incorporate non-response as a partitioning criterion.

These results indicate substantive effects between the future student aspirations (those who intended to attain a baccalaureate degree) and overall performance, with these students performing significantly better on all three sections. The failure for the function to further discretize the low achievers indicates that the individual differences within this group accounted for no additional information. For high aspiring students, ethnicity was found to be the second most important variable
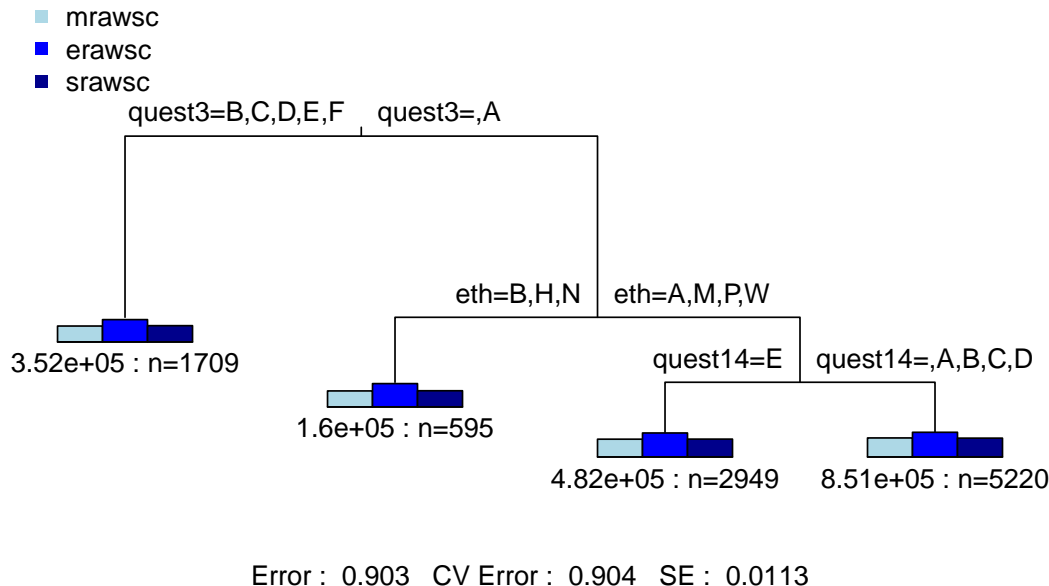
mrawsc
erawsc
srawsc

quest3=B,C,D,E,F   quest3=,A

3.52e+05 : n=1709

eth=B,H,N | eth=A,M,P,W

1.6e+05 : n=595

quest14=E | quest14=,A,B,C,D

4.82e+05 : n=2949   8.51e+05 : n=5220

Error : 0.903   CV Error : 0.904   SE : 0.0113

Figure 1: Multivariate regression tree of questionnaire results. *'mrawsc'*, *'erawsc'*, and *'srawc'* are the total Mathematics, English, and Biology scores for each student, respectively. *'questX'* is questionnaire question X, and *'eth'* is the reported ethnicity.

in discriminating test scores: Blacks, Hispanics and Native Americans performed similarly worse on the Mathematics and Biology sections compared to the other ethnicities. The final node split discriminates between high and low scoring science students, with those who responded as rarely using scientific instruments performing worse. Cross-validated errors were high, at CVE=.905, which corresponds to R^2=.15.

4.6.2   Multivariate Regression

A parametric regression model was also conducted exploring only main effects between the predictors and response. Secondary analyses run upon the data, using the VIM package, clearly indicated that the missing data was uniformly equal for all but one question, which substantively improves the likelihood that these questions are truly missing completely at random (MCAR). In addition, the bulk of the missing information was a result of all questions being skipped together. Unfortunately, more algorithmically driven tests (Little, 1988; Kim & Bentler, 2002), traditionally require that the data in question be normally distributed, and then proceed with homoscedatic assessments of the data; for our purposes, the predictors are all categorical. All tests results were based on both the imputed and missing data, however no differences were found amongst the missing results. Results were singly focused upon the science sections, with more involved educational settings supporting higher outcomes. Males typically outperformed women, and Hispanic/Latino and Black students performed significantly worse than other ethnicity groups, who performed non-significantly differ-

entially from themselves. Otherwise, student achievement outcomes were the only conclusively meaningful predictors that presented, with students who expected to achieve higher (e.g., achieve a baccalaureate degree) having higher outcomes. All code necessary to replicate these effects are submitted within the R script file submitted with these results.

The regression tree can capture interactions between variables and nonlinear effects, while the multivariate regression performed only measures linear effects. The fact that the R^2 values for each analysis are similar suggests that there may not be significant interactions or nonlinear effects in the relationship between questionnaire results and test scores.

5   Discussion

The primary point of order in test design is for a test to perform equivalently across groups of testees. As indicated in the results section, our analysis suggests that the test fails to perform equivalently across four ethnicity groups. This supplies quantitative evidence to the argument that there are fundamental flaws within the test. However, the multivariate regression and tree results indicate that there are no functional differences between the test scores themselves across ethnicity, which is surprising given the previous point. In our opinion, a plausible explanation for these results are that there are ethnic differences not between the overall scores but between general education practices used for each group, likely a function of SES, which disproportionately affects minorities. This subtle distinction lends credence to the argument that there is a fundamental flaw within the standards of difficulty construed as appropriate for this test, and as a consequence, the test performs poorly as an indicator of individual student achievement.

Little more can be said about the questionnaire beyond the fact that the mode for each item is the most efficient predictor of individual student responses. As such, we can only conclude that the information being measured by these items is independent of the exams' item functioning and student performance.

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

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

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, *83*(4), 1105–1117.

Gaudet, R. D. (2000). Effective school districts in massachusetts. *The Donahue Institute, Boston*.

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*(4), 609–623.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.

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

Pornprasertmanit, S., Miller, P., Schoemann, A., & Rosseel, Y. (2013). semtools: Useful tools for structural equation modeling. [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=semTools` (R package version 0.4-0)
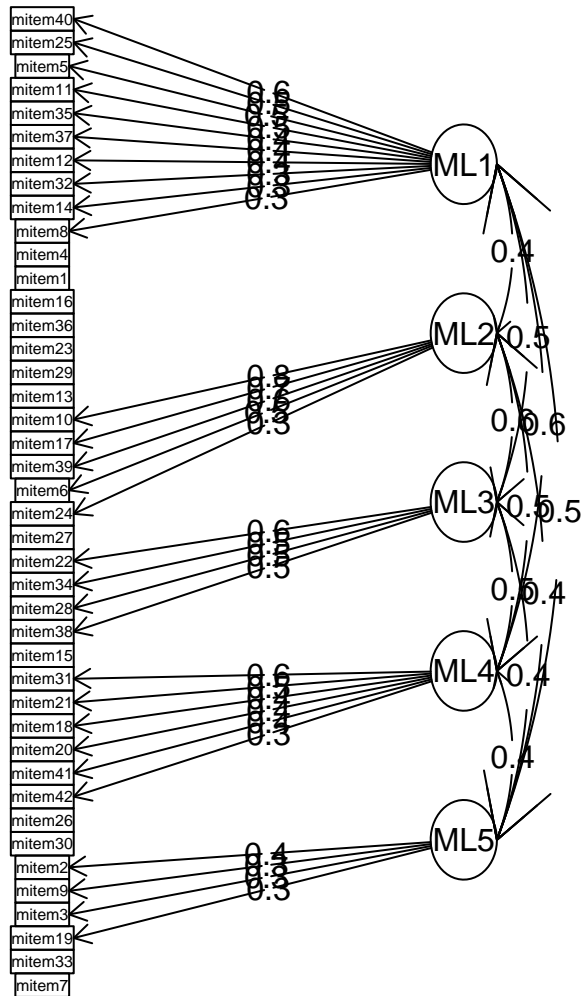
Revelle, W.   (2014).   psych: Procedures for psychological, psychometric, and per-
        sonality research [Computer software manual].   Evanston, Illinois.   Retrieved from
        `http://CRAN.R-project.org/package=psych` (R package version 1.4.2)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical
        Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys* (Vol. 307). John Wiley &
        Sons.

Spring 2009 mcas tests: Summary of state results. (2009, September).

van Buuren, S., & Groothuis-Oudshoorn, K.   (2011).   mice: Multivariate imputation by
        chained equations in r.   *Journal of Statistical Software*, *45*(3), 1–67.   Retrieved from
        `http://www.jstatsoft.org/v45/i03/`

6 Appendix A: Exploratory Factor Analysis

English Oblimin
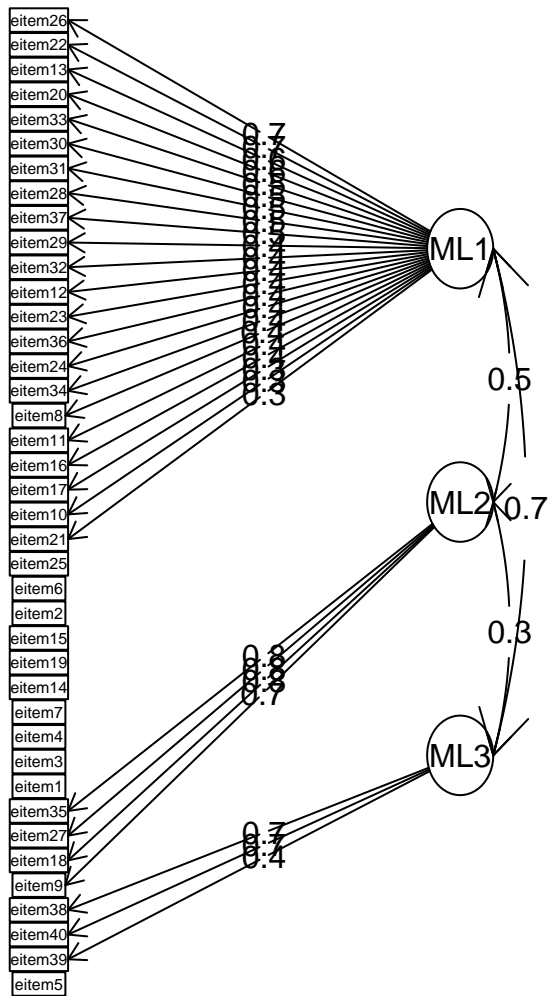
Mathematics Oblimin

Biology Oblimin