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

2 Introduction

2.1 Data

The Massachusetts Compreshensive 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 10^{th} 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 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. 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 unrealitic 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 analaysis of both the psychometric equivalance of the MCAS test within our sample of 10^{th} grade students, followed by an investigration into performance within each knowl-

edge domain, conditioned upon the demographics and item responses found within a questionnaire included with the test.

3 Methods

3.1 Subjects

Our sample is comprised of 10,515 10^{th} grade students enrolled in Massachusett's public education system, and is assumed to be drawn as an unbiased sample of the total universe of 10^{th} 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 relaionship with student scores (Gaudet, 2000), as mediated by ethnicity or spatial location.

3.2 Procedure

A critical assumption within standardised 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 underly the observed measures; one typically estiamtes 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, 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 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 preceeding 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 group of R packages cited at the end of the paper (Revelle, 2014; Rosseel, 2012; Pornprasertmanit, Miller, Schoemann, & Rosseel, 2013). Exploratory bootstrapped oblimin factor analyses were conducted upon each section 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.

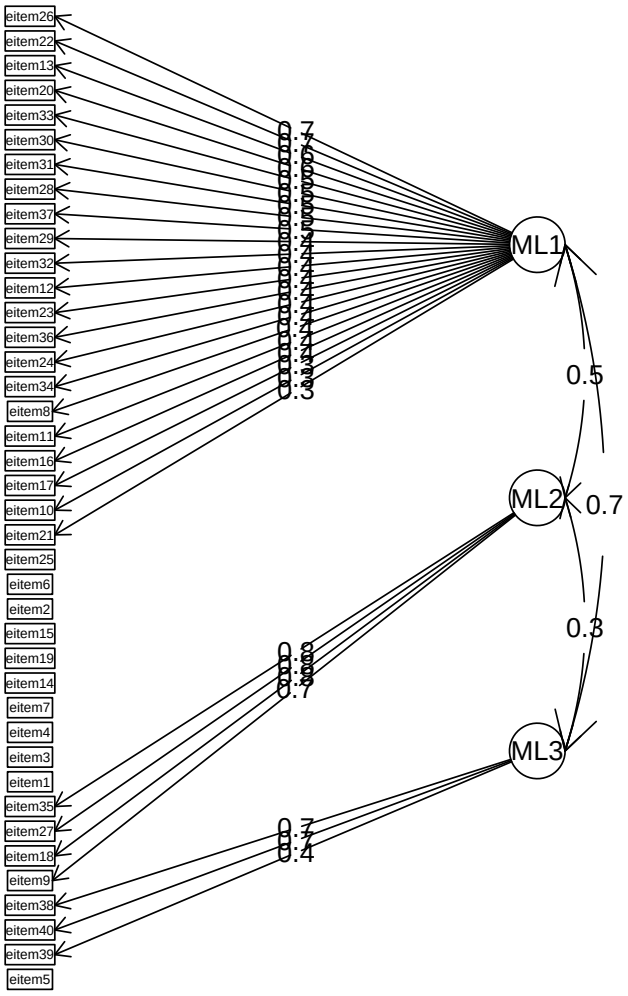
Furthermore, we investigated the information found within the questionnaire, relating them to the students' scores on the three sections using both multivariate parametric regressions, and a non-parametric predictive procedure using the mvpart package, which models multivariate regression trees.

Missing data in survey and test data is an important consideration, and there exist a number of techniques to produce more robust solutions, mainly working 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 occurence: 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 utilising 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, producing potentially biased results under the assumption that items are missing completely at random. As such, we processed the original responses using the mice package(Buuren & Groothuis-Oudshoorn, 2011), using 25 imputed data frames with 30 iterations in each dataframe. Regression trees are relatively robust to missing data, because all information is not processed simultaneously, but sequentially in order of variable importance, and as such we did not take any additional steps when conducting that analysis.

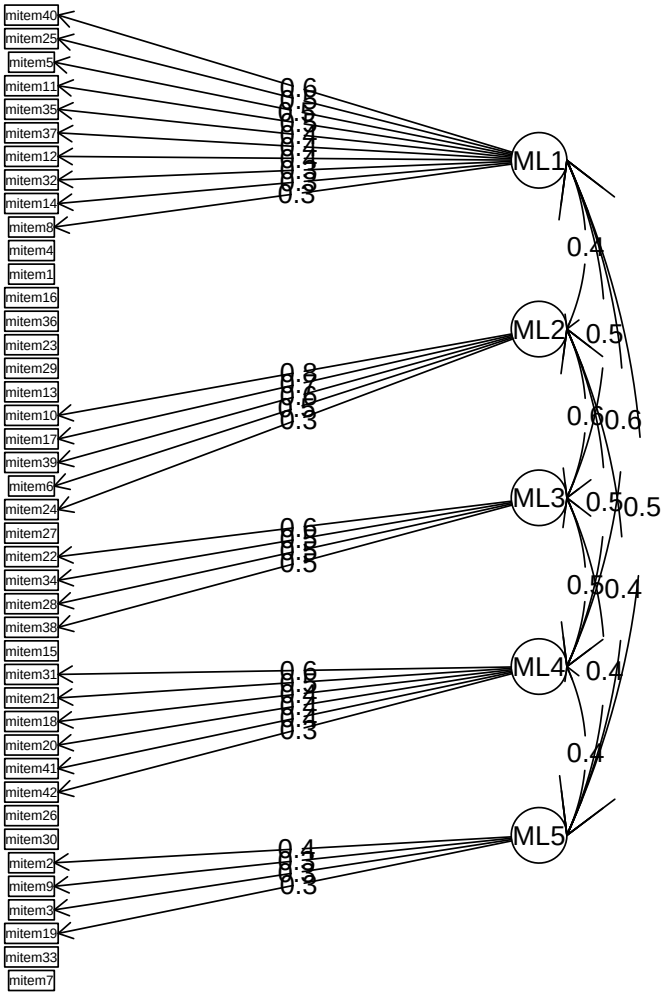
4 Results

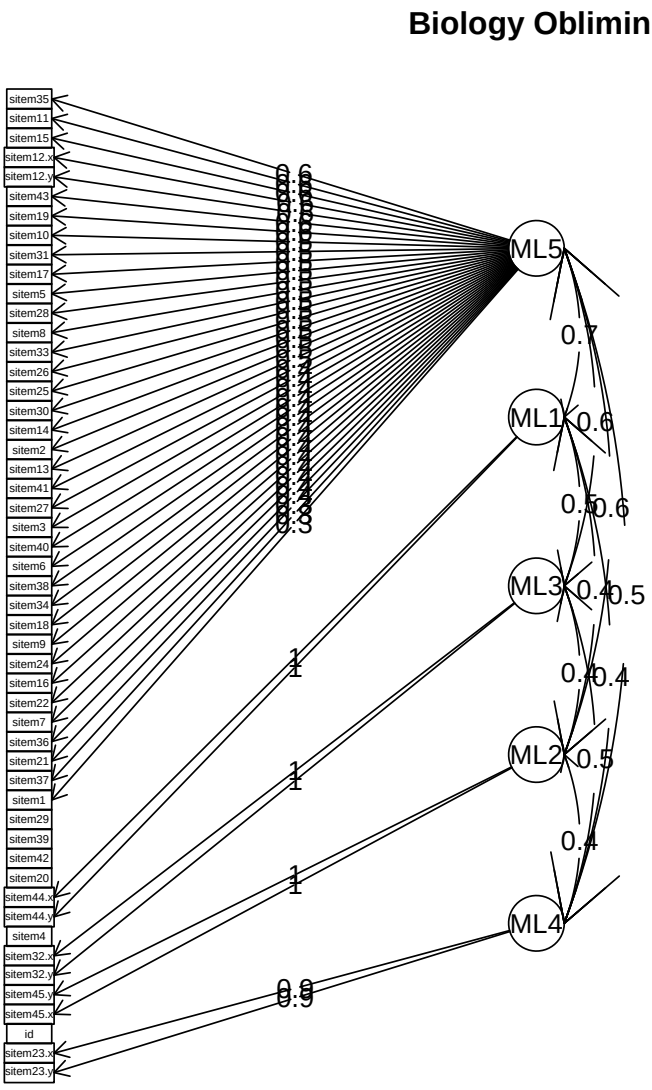
4.1 Exploratory factor analyses

Given the preceeding 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.


Mathematics Oblimin
mitem40
mitem25
mitem5
mitem11
mitem35
mitem37
mitem12
mitem32
mitem14
mitem8
mitem4
mitem1
mitem16
mitem36
mitem23
mitem29
mitem13
mitem10
mitem17
mitem39
mitem6
mitem24
mitem27
mitem22
mitem34
mitem28
mitem38
mitem15
mitem31
mitem21
mitem18
mitem20
mitem41
mitem42
mitem26
mitem30
mitem2
mitem9
mitem3
mitem19
mitem33
mitem7
ML1
ML2
ML3
ML4
ML5
0.4
0.5
0.6
0.6
0.5
0.5
0.4
0.4
0.4



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 three separate tests for measurement invariance:

These were subsequently 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. As a consequence of the χ^2 fit statistic, defined as

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

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 constructed using the following system of latent and test items, as predictors of the unstandardized English score, as given below.

```
## erawsc ~ e1 + e2 + e3
```

4.3.1 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.

4.3.2 Metric (Weak) Invariance

Weak Invariance does impact the model fit compared to step 1, however negligably: Δ CFI = .002, Δ RMSEA = .000 , $\Delta\chi^2$=187.875, and the Δ TLI =.

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

4.3.3 Strong Invariance

Strong Invariance is established using the model fit indexes above and beyond weak invariance: Δ CFI = .002, Δ RMSEA = .000 , and the Δ TLI =.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
```

4.3.4 Strict Invariance

Strict invariance does not hold, with a significant Δ CFI = .034, Δ RMSEA = .006, $\Delta\chi^2$=1670.297, and the Δ TLI = .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 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.

```
## mrawsc ~ m1 + m2 + m3 + m4
```

4.4.1 Configural Invariance

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

4.4.2 Metric (Weak) Invariance

Weak Invariance does impact the model fit compared to step 1, however negligably: Δ CFI = .004, Δ RMSEA = .001 , $\Delta\chi^2$=260.076, and the Δ TLI = .002.

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

4.4.3 Strong Invariance

Strong Invariance does not substantively change model fit above and beyond weak invariance: Δ CFI = .002, Δ RMSEA = .000, the Δ TLI = .000 and $\Delta\chi^2$=167.652.

```
##     chisq       tli       cfi     rmsea       aic
## 3.132e+03 9.630e-01 9.630e-01 2.800e-02 3.162e+05
```

4.4.4 Strict Invariance

Strong Invariance imposes a significant constraint upon model fit: Δ CFI = .017, Δ RMSEA = .005, the Δ TLI = .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 as presented an interpretable factor structure, which was represented with the following mathematical structure.

```
## srawsc ~ s1 + s2 + s3 + s4
```

4.5.1 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.

4.5.2 Metric (Weak) Invariance

Weak Invariance does impact the model fit compared to step 1, however negligably: Δ CFI = .002, Δ RMSEA = .000 , $\Delta\chi^2$=168.330, and the Δ TLI = .000.

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

4.5.3 Strong Invariance

Strong Invariance does fails to substantively change model fit above and beyond metric invariance: Δ CFI = .002, Δ RMSEA = .000 , and the Δ TLI = .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
```

4.5.4 Strict Invariance

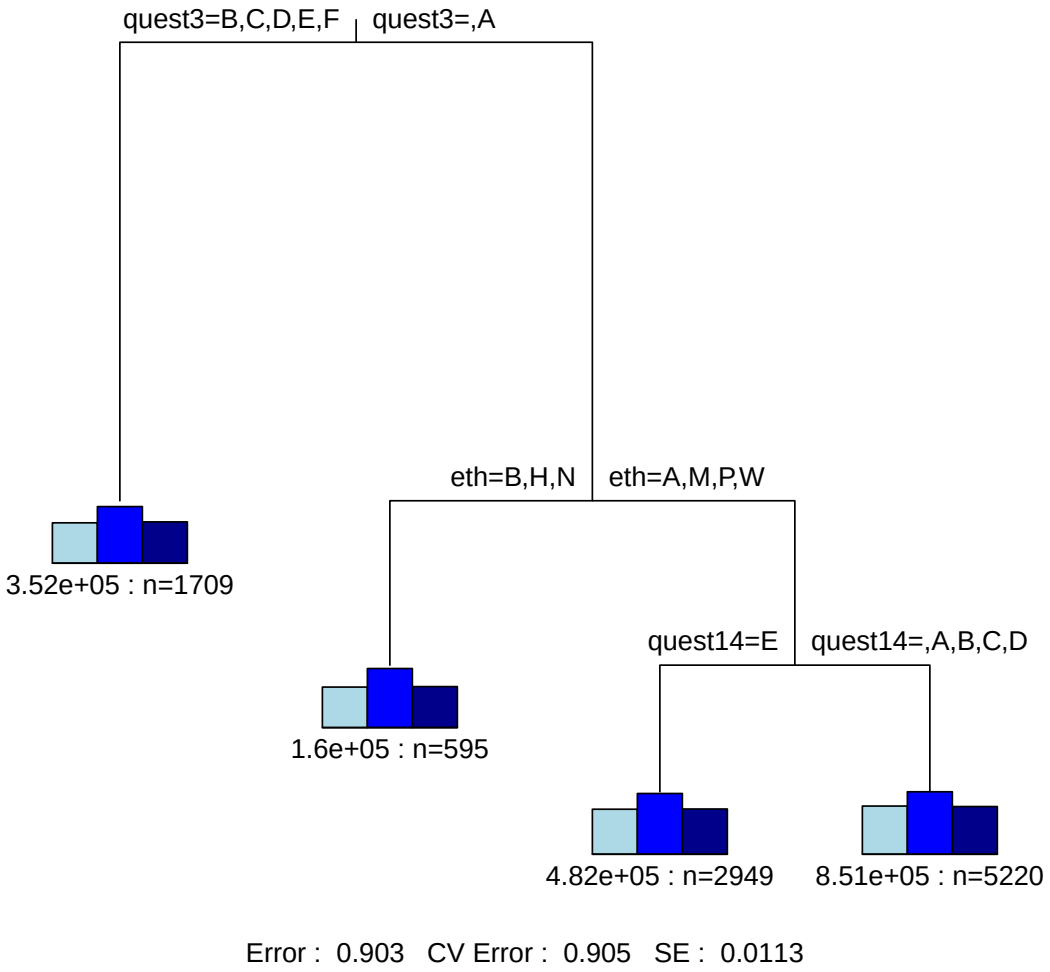
Strict Invariance does substantively change model fit above and beyond strong invariance: Δ CFI = .034, Δ RMSEA = .006 , and the Δ TLI = .03 and $\Delta\chi^2$=1670.297.

```
##     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), conceptually similar to running a multinomial regression upon the different clusters analysis groups 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 nonresponse 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 wihin accounted for no additional information. For high aspiring students, ethnicity was found to be the second most important in discriminating test scores, with Blacks, Hispanics and Native Americans performing similarly worse on the Mathematics and Biology sections than 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.

5 Discussion

As indicated in the results section, the test fails to perform equivalently across the four groups, which is the primary point of order in test design and advocacy. Moreover, strict invariance has a specific import upon the function of psychometric and education measurements, testing whether the error variances –the portions of item variation not attributable to the common factors. Conceptually, strict invariance validates the assumption that error variances are different across groups; if this fails, there are either different variables operating on the measures across groups or the same set of variables operates differently across groups (DeShon, 2004). This fundamentally demonstrates the importance of ensuring that all four mainstays of measurement invariance be substantiated, especially when one considers the long standing view that residuals convey meaning of unmeasured latent constructs' effects, moderated by group membership in this case (Cronbach, 1947). This lends substantive credence to the argument that there are fundamental flaws within the test, specifically with reflection towards students' socioeconomic status.

References

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

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

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

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, *12*(1), 1–16.

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

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137–149.

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

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]. Available from `http://CRAN.R-project.org/package=semTools` (R package version 0.4-0)

Revelle, W. (2014). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Available 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. Available 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).