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 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. 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.

Our analysis 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 knowledge 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} grade ethnicity characteristics as 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.

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 is that the test demonstrates a psychometric property called measurement invariance/equivalence. This property is a series of mathematical constraints applied to latent variables. Latent variables are unobservable constructs measured in an identical process to 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 iterative four phase procedure:

- Configural invariance: Tests whether the same factor model (i.e., latent variables, equivalent to knowledge subdomains, for example grammar) is found within each subgroup.
- Weak invariance: Tests whether the same items load equivalently onto the same same structure, presuming that configural invariance has been demonstrated.
- Strong invariance Tests that the intercepts are equivalent across groups.
- Strict invariance: Tests the assumption that 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 found 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 Theoretical structure of the test

Substantive theoretical knowledge of the number of latent knowledge domains that were intended to be tested was generated from the answer keys provided by Massachusetts' Department of Education. Specifically, there were 5 general topics within Biology:

- Anatomy & Physiology
- Biochemistry and Cell Biology
- Ecology
- Evolution and Biodiversity
- Genetics

Two topics within English:

- Reading & Literature
- Language

and five topics within Mathematics:

- Data Analysis Statistics and Probability
- Geometry
- Measurement
- Number Sense and Operations
- Patterns Relations and Algebra

4.2 Exploratory factor analyses

With 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, and factor loadings calculated using maximum likelihood estimates calculated using the tetrachoric covariance matrices for each subset of items. The psych package was utilised both 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 B.

4.3 MG-CFA

As a consequence of the unequal numbers of test takers between Biology and the rest of the exam, we chose to run two separate tests for measurement invariance by running Biology separately. Using the results of the EFA, we specified the following two models:

```
invariance
```

```
## Error:  object 'invariance' not found
```

```
sci_inv
```

```
## Error:  object 'sci_inv' not found
```

These were subsequently specified using a FIML procedure with bootstrapped sampling within the four primary ethnicities: A,B,H,W. Others were excluded due to sample size and composition constraints. As a consequence of the χ^2 fit statistic scaling as a function of both sample size and the minimum function test statistic, traditional measurement invariance techniques of using changes in significance to reject a constraint fail here, as the sample size guarantees significant results. In response to this fact, we employ two different techniques: 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 both, decreases in model fit of greater than .01 are considered significant. In addition, we include An (Akaike's) Information Criterion (AIC), to demonstrate whether model fit increases or decreases between sequential steps.

4.4 English and Mathematics

4.4.1 Configural Invariance

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.

```
fitMeasures(MI.model$fit.configural, c("chisq", "tli", "cfi", "rmsea", "aic"))
```

```
## Error:  could not find function "fitMeasures"
```

4.4.2 Metric Invariance

```
fitMeasures(MI.model$fit.loadings, c("chisq", "tli", "cfi", "rmsea", "aic"))
```

```
## Error:  could not find function "fitMeasures"
```

The second test demonstrates a failure of metric/weak invariance when comparing either the Δ TLI or Δ CFI. Commonly, metric invariance is used to demonstrate that for all items, one unit change in the item scores is scaled to an equal unit change in the factor score across groups. This implies that the variances on each measure are the same for all items. As a consequence of the failure of metric invarianc to hold, all subsequent equivalences are rejected as well.

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.

```
fitMeasures(MI.model2$fit.configural, c("chisq", "tli", "cfi", "rmsea", "aic"))

## Error:  could not find function "fitMeasures"

fitMeasures(MI.model2$fit.loadings, c("chisq", "tli", "cfi", "rmsea", "aic"))

## Error:  could not find function "fitMeasures"

fitMeasures(MI.model2$fit.intercepts, c("chisq", "tli", "cfi", "rmsea", "aic"))

## Error:  could not find function "fitMeasures"

fitMeasures(MI.model2$fit.means, c("chisq", "tli", "cfi", "rmsea", "aic"))

## Error:  could not find function "fitMeasures"
```

4.6 Questionnaire

Preliminary results were conducted using a multivariate regression tree, which searches for both linear and non-linear relationships between different classifications of the three raw scores and the questionnaire items, ethncity, and gender. However, the data failed to split at all, meaning that there is little information that relates between the questionnaire, demographics, and individuals' scores. Likewise mutlivariate regression results included demonstrate the no meaningful results were identified, with the few significant results conceptually meaningless.

```
require(mvpart)

## Loading required package:  mvpart

fit1 <- mvpart(cbind(d$mrawsc, d$erawsc, d$srawsc) ~ ., data = d4)

## Error:  object 'd4' not found

summary(manova(cbind(d$mrawsc, d$erawsc, d$srawsc) ~ ., d4))

## Error:  object 'd4' not found
```

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. This lends substantive credence 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 surpising 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 the general education practices taught, which is likely a function of SES, which disproportionately affects minorities over Whites. This subtle distinction has the immediate consequence of arguing 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 are 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.

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.

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).