Final Project

Reevaluating the "Mechanisms and Impacts of Gender Peer Effects at School"

Elizaveta Gonchar

PSYC 8060 - Georgia Institute of Technology

December 2019

1 Background and Motivation

For this project, I thought it would be interesting to consider a paper from a top journal in economics and reconsider the authors' findings using item response theory (IRT) methods. When I came across Lavy & Schlosser (2011b)[1], I realized that the data used in the economic analysis lends itself to IRT methods because the justification used in the analysis mirrors that of those used in IRT. Using the Growth and Effectiveness Measures for Schools (GEMS) data and academic achievement measures for students of all public schools in Israel, the authors identified channels through which the presence of girls in the classroom via gender peer effects lead to improvements in academic achievements. The intent of their paper was to estimate the effect the creation of single-sex classes has via the change in the sex ratio of coeducational classes. Israel is a nearly-ideal case study to examine these possible effects, due to low inter-school mobility stemming from lack of choice, scarce private schooling options, the significant coverage of the existing school surveys, and the fact that the results of the surveys are not publicly available. The characteristics of the educational structure in Israel contributed to the identification of gender peer effects, as schooling choice is unlikely to be endogenous, particularly on the basis of classroom gender composition or survey results. The longitudinal nature of the data also allowed the authors "to examine the impacts of changes in peers' gender composition within the same student" (Lavy & Schlosser, 2011b, p. 4). While this portion of the analysis was intriguing, I do not make use of this characteristic of the data in my analysis.

The authors considered and presented multiple mechanisms and channels to show that the proportion of girls in a classroom as well as the increase in this proportion is strongly correlated with improvements in student achievement (captured by test scores) of the respective cohort for elementary, middle, and high school students in Israel. Based on the questionnaire and test score data from public school in Israel, the author's identified the following mechanisms through which changes in the gender composition within a school affect the scholastic outcomes of students: disruption and violence, inter-student relationships, teacher-student relationships,

[1]Note that the authors published a corrigendum (Lavy & Schlosser, 2011a).

self-discipline, and study efforts. Lavy & Schlosser (2011b) conclude that "an increase in the proportion of girls improves boys and girls' cognitive outcomes" (Lavy & Schlosser, 2011b, p. 1); additionally, the authors conclude that this effect is channeled via changes in the classroom environment as opposed to changes in students' individual behaviors.

The goal of my research is to consider how the use of IRT methods, particularly bifactor estimation, will support or contradict the basis of Lavy & Schlosser (2011b), an econometrically driven paper. I use the data provided by the authors in their replication kit to conduct my IRT estimations. Since is the most comprehensive, I focus on the elementary school data for my analysis[2]. In Section 2, I provide a comprehensive description of the data available for the elementary school students.

For my analysis, I conducted three estimations using the graded response model: unidimensional, bifactor, and bifactor with subgroups. The main estimation strategy used for this project was a bifactor model approach which I used with the intent of verifying the factor loadings as identified by the authors. I provide details on my estimation strategies in Section 3. I find that the items do not load on to the factors as indicated by the authors. Based on the three estimations I conducted for 5^{th} grade students, I find that there is reason to believe that the structure of the items as identified by the authors is not necessarily accurate.

2 Data

I use the data set provided by Lavy & Schlosser (2011c) for replication purposes. I rely heavily on the summary provided by the authors when describing the data used in my analysis. For their analysis, the authors used test score data for elementary, middle, and high school students along with survey data, which is only available for elementary and middle school students. My intent is to conduct a confirmatory analysis regarding the authors' proposed factor loadings; for this reason, I will be working with a sub-sample of the survey data provided in the replication kit by focusing on 5^{th} grade students. I will first provide a brief overview of what GEMS is and will then provide an overview of the data available in the replication kit.

[2]GEMS is only administrated at the elementary and middle school levels.

The GEMS survey is administered annually by the Division of Evaluation and Measurement of the Ministry of Education in Israel. According to Justman & Bukobza (2010), the survey is intended to measure the scholastic achievement of students, as well as to capture the climate of their classrooms and schools. All of the survey items are on a 6-point scale items, ranging from 1 (strongly agree) to 6 (strongly disagree) for the extent to which they agree with a series of statements. The survey also asks students to report average hours (weekly) spent on homework in math, Hebrew, English, and science and technology; the responses found in the data range between 0 and 5 hours for all 4 subjects. GEMS is unique and provides an interesting basis of study because each school is usually sampled every 2 years[3] with a completion rate of approximately 91%. Additionally the National Authority for Measurement and Evaluation in Education issues reports on the results for the Ministry of Education's internal use and the public and parents have no access to reports on the results.

For elementary and middle school students, the authors obtained GEMS survey and test score data; this data was linked to the students' administrative records which include student background characteristics and demographics. For this analysis we have two observations of the same school and grade for more than 90% of the schools, measured from 2002 to 2005. Additionally, this data contains information on the proportion of girls in each cohort for each school. I provide descriptive information regarding the total data in the proportion of females in each school cohort which can be found in the appendix (Section A). We can see that the proportion follows a relatively normal distribution.

For elementary school, the authors had test school and questionnaire data for 5^{th} grade students; they had questionnaire data for 6^{th} grade students as well but the lack of test score data resulted in the exclusion of these students from the analysis. The GEMS data for 5^{th} grade students covers 1,010 elementary schools (808 secular and 202 religious) and the test score data covers 997 elementary schools (808 secular and 189 religious). Given the time frame (2000 to 2005), the data contained two observations of test scores and questionnaires for each

[3]Each year a school has a 50% chance of being surveyed. The target is to have schools surveyed every other year and this is generally true.

elementary school.

For middle school, the authors had test school and questionnaire data for 8^{th} grade students; they had questionnaire data for 7^{th} through 10^{th} grade students as well but the lack of test score data resulted in the exclusion of these students from the analysis. The sample of schools used only included secular schools "since there are only a few religious middle schools with mixed-gender classes" (Lavy & Schlosser, 2011b, p. 9). This left us with 395 secular schools in the sample, of which 85% appear in two years.

For high school students, they acquired administrative records which were collected by the Israel Ministry of Education for 8 consecutive cohorts of 10^{th} graders from 1993 to 2000. This data contains an individual as well as class identifier along with detailed demographic information on each student. This administrative data was linked with the following matriculation outcomes: average score in the matriculation exams, matriculation status, number of credit units, number of advanced level subjects in science, and matriculation status that meets university entrance requirements. The reason the authors only consider 10^{th} grade students is that it's the first year of high school and the last year of mandatory schooling in Israel.

As I began working with the data in IRTPRO, I realized binding constraints of time and also a lack of standardization of knowledge across years across schools; It would be optimal to hone my estimation in on one year. I decided to consider the survey data for 5^{th} and 8^{th} grade students from 2003. This left me with 27,281 observations for 5^{th} grade students and 24,189 for 8^{th} grade students. After presenting my initial results on December 6, I determined that it would behoove me to focus my attention on 5^{th} grade students only because there are more schools covered by the data and it seems that the distribution of proportion of girls in a class is wider[4]. For my analysis on 5^{th} grade student survey response data, I was left with 27,281 observations with complete data for 23,267 of those observations.

In Figure 2, I present the traditional summary statistics of the data and I provide the complete output in Section B. There are a few features that I felt necessary to note. First,

[4]I provide summary statistics as well as a histogram of the proportion of female students in a 5^{th} grade cohort in Section A.

the items all have fairly similar Cronbach's α values with a total coefficient α of 0.68, which is fairly low. Second, items 4, 9, and 10 (*q41R*, *q31R*, and *q32R*, respectively) have observations that are heavily favoring one side of the Likert scale; this may be concerning, particularly for *q32R* which only had 62 observations for a response of 2 (since item is reverse coded). Overall, the data is certainly not without its flaws but it is able to overcome identification issues as a result of the large number of observations.

3 Methods

For my analysis, I conducted three estimations using the graded response model: unidimensional, bifactor, and bifactor with subgroups. I decided to implement the graded response model because the items used had ordered response categories and the number of response categories was equal for all of the items. I conducted unidimensional analysis as a baseline model to provide a form of comparison for the bifactor model estimation. The main estimation strategy used for this project was a bifactor model approach which I used with the intent of verifying the factor loadings as identified by the authors. The extension I chose to implement was suggested to me by Dr. Embretson: group school cohorts by proportion of girls (low, moderate, and high) and determine whether the results differ across groups. This estimation was motivated further by the heterogeneous effects identified by the authors. In this paper, I present the estimation results of a proposed bi-factor design (presented in Figure 1) that I believe captures the intent of Lavy & Schlosser. I provide an interpretation of the model in the section that follows. As an extension to motivate the considerations of the authors, I expand the bifactor analysis to incorporate subgroups. Based on the reported gender proportions available for each observation, I divide the data into three groups: low female-in-class proportion, moderate female-in-class proportion, and high female-in-class proportion. Following a standard used in economics, I identified low-proportion observations as those with a female-in-class proportion at or below the 25th percentile of the data. Similarly, I identified high-proportion observations as those with a female-in-class proportion at or above the 75th percentile of the data. Those

with a female-in-class proportion between the 25th and 75th percentile were identified to belong in the moderate proportion group. This subgroup estimation is conducted to complement the estimations the authors conducted on the gains from having females in a classroom. The idea being that individuals in classrooms with greater proportions of females are likely to have higher estimated traits for classroom environment while seeing very little variation with regards to students' behavior items. I continue by providing an overview of how I implemented this proposed measurement application.

I began by conducting a baseline estimation using the graded response model. I ran the unidimensional estimation for all the observations (2002-2005) for 5^{th} and 8^{th} grade with 106,119 and 93,442 observations, respectively. These results are not provided in this paper. As I began to implement the bifactor model estimation on the data, I realized that it would not be feasible or necessarily insightful to conduct the estimation across the four years. As I mentioned in Section 2, I decided to cut down my sample to only include observations from 2003 focusing only on 5^{th} grade students. The results of the unidimensional estimations of the 5^{th} grade students are presented in Section 7. After my presentation, I decided to focus my considerations on 5^{th} grade students from 2003; this decision was motivated primarily to allow me to focus on one complete set of data interpretations without having to incorporate considerations across grades. I feel that the decision to exclude 8^{th} grade students from my analysis is validated by the fact that the estimations I was obtaining were fairly similar to those of the 5^{th} grade students.

The crux of the Lavy & Schlosser (2011b) paper analysis relies on the validity of their survey item groupings. The authors do not appear to provide direct justifications for the categorizations of the survey items that they implemented. For example, with regards to the items grouped under inter-student relationships, the authors provide an intuitive justification to explain why these items belong in the category; while we cannot rule out the possibility that the factors were estimated using the appropriate methods, the lack of evidence presented on the matter motivated me to pursue this estimation.

As mentioned earlier, given the proposed factors identified by the authors, I deemed it

appropriate to conduct a multidimensional confirmatory analysis on the data. To do so, I constructed a bi-factor estimation model containing three latent traits. The first trait is the general trait which, per the authors' rationale, should be achievement relevancy as the items have been identified to be related to academic outcomes of students. The second trait is the classroom environment which is believed to be captured by the eight items relating to disruption and violence, inter-student relationships, and teacher-student relationships. The third trait is the students' behavior which is believed to be captured by self-discipline[5] and study efforts. If the authors have correctly identified the appropriate latent factors, the bi-factor model results should indicate such via the estimated factor loadings. The results of the estimation are presented in Section 7. Finally I implemented the subgroup bifactor estimation as discussed above; these results are presented in Section 7 as well.

4 Results

As mentioned earlier, the IRTPRO outputs have been provided in Section 7. I provide the factor loadings, goodness of fit, and item parameter estimates for all three estimation techniques for 5^{th} grade students. The results of the bifactor model carry over the the subgroup bifactor estimation; thus I do not feel it is necessary to discuss the results of the estimation in detail as it did not seem to enhance the analysis. I present the goodness of fit statistics of -2logLikelihood in Figure 3. I find that estimated χ^2 of the bifactor model is 42,000.23 which indicates that the model fit improves with the bifactor estimation.

I will begin by discussing the factor loading results, presented in Figures 4 and 5, across the three estimations as they are fairly similar. If we consider items *q47* to *q50*, which are the average hours spent on homework for a given subject per week, we see that it loads strongly in all three estimations on to the students' behavior trait; however, it does not load highly on to the achievement relevancy trait which is likely to indicate that time spent on homework does not directly correspond to academic outcomes. If we consider the remaining items of the

[5]Note that self-discipline in this case means "the student's understanding of the learning and discipline requirements in school, his/her involvement in fights with other students, and his/her relationship with the teachers" (Lavy & Schlosser, 2011b, p. 29)

student behavior trait (*q31*, *q32*, *q38*, *q40*, and *q45*), we see that these items do not load onto the student behavior latent trait in any meaningful way and we may say that items *q31R* and *q32R* are slightly overestimated on the student behavior latent trait.

Let's consider the items that are identified as loading on to classroom environment. The classroom disruption and violence items (*q34*, *q37*, and *q39*) seem to load fairly strongly onto the environment trait as well as the achievement relevancy factor though not as strongly. The inter-student relationship items (*q41R* and *q42R*) do not load on to the classroom environment factor at all but do load strongly onto the achievement relevancy general factor. Finally, all three items *q35*, *q43R*, and *q44R*, corresponding to teacher-student relationships, load fairly strongly onto the achievement relevancy factor; however only *q35* loads onto the classroom environment factor and the remaining items in teacher-student relationships (*q43R* and *q44R*) are overestimated on the classroom environment trait. These results seem to indicate that the existence of these two latent traits as the authors have defined them is not necessarily valid. This conclusion seems to be corroborated by the item parameter estimates, presented in Figures 6 and 7. Overall, while there are some items load strongly to their unique latent trait, I am led to question the structure as defined by the authors given the fact that there are many items that are strongly overestimated in their structure.

5 Discussion

The most notable limitation in my analysis was the data available to me. For this project, I used the replication data made publicly available by the authors[6]. This data was fairly comprehensive and contained most[7] of the relevant data; however, the lack of data on the survey items that were not included in the analysis hindered the possibility of a exploratory multidimensional analysis. If I were to continue with this project, I would be interested in acquiring, at minimum, a complete list of the questionnaire items.

As I was working on this project, I thought it would behoove me to research GEMS. I came

[6]Available on the American Economic Association article page as well as ICPSR.

[7]The authors do not provide the data containing teacher survey responses, which, fortunately, was not necessary for this analysis.

across a report (Justman & Bukobza, 2010) that presented the intent of the survey along with its shortcomings; this provided me with a better sense of where the issues with this estimation method would be. In particular, it was acknowledged that, due to a lack of set standards regarding academic attainment by grade level, it is difficult to make comparisons across schools with regards to test scores. While this does not necessarily hinder the research conducted here, it is important to bear this limitation in mind.

Prior to making any determinations regarding the structure of the survey items, I believe an exploratory multidimensionial factor analysis is necessary. While I do not believe the latent traits identified by the authors appear as they have indicated, I do believe that there are multiple channels through which the academic achievements of students are affected. Though I was not able to provide an definitive evidence or conclusion in this paper, I believe the results presented here lend themselves to guiding future research in this area.

References

Justman, M., & Bukobza, G. (Eds.). (2010). Guidelines to Revise the System of Education Indicators in Israel: Summary, Conclusions and Recommendations, By the Committee to Revise the System of Education Indicators in Israel. Jerusalem: Israel Academy of Sciences and Humanities.

Lavy, V., & Schlosser, A. (2011a, July). Corrigendum: Mechanisms and Impacts of Gender Peer Effects at School. American Economic Journal: Applied Economics, 3(3), 268-268. Retrieved from http://www.aeaweb.org/articles?id=10.1257/app.3.3.268 doi: 10.1257/app.3.3.268

Lavy, V., & Schlosser, A. (2011b, April). Mechanisms and Impacts of Gender Peer Effects at School. American Economic Journal: Applied Economics, 3(2), 1-33. Retrieved from http://www.aeaweb.org/articles?id=10.1257/app.3.2.1 doi: 10.1257/app.3.2.1

Lavy, V., & Schlosser, A. (2011c). Replication data for: Mechanisms and Impacts of Gender Peer Effects at School. Nashville, TN: American Economic Association. Retrieved from https://doi.org/10.3886/E113785V1 doi: 10.3886/E113785V1

6 Figures

Figure 1: Proposed Bifactor Design

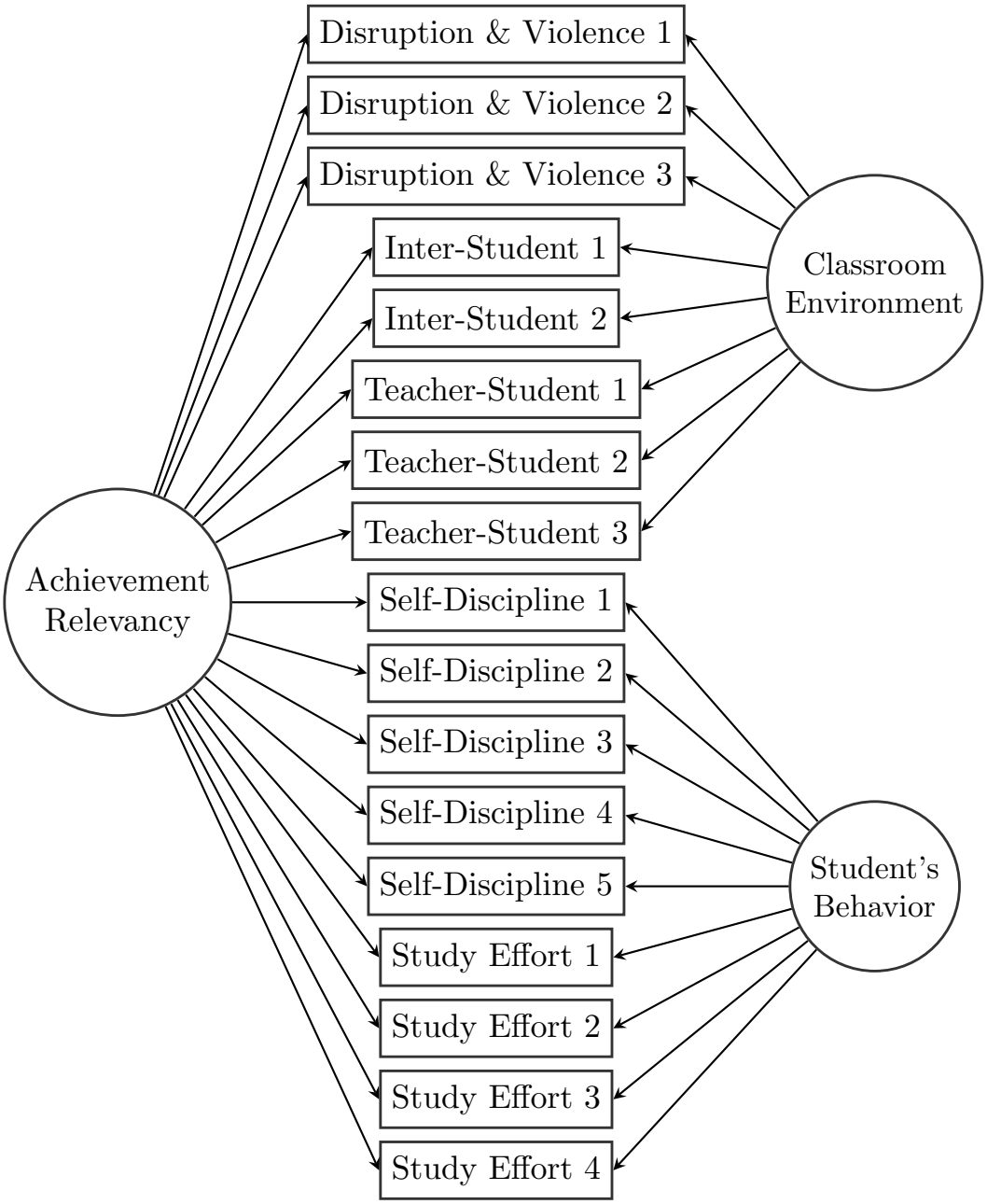


Table 1: Overview of Survey Items

Classroom Environment	
Classroom Disruption and Violence	
q34	(1) Frequently the classroom is noisy and not conducive to learning.
q37	(2) There are many fights among students in my classroom.
q36	(3) Sometimes I'm scared to go to school because there are violent students.
Inter-Student Relationships	
q41R	(4) I feel well adjusted socially in my class.
q42R	(5) Students in my class help each other.
Teacher-Student Relationships	
q35	(6) Students frequently talk back to teachers.
q43R	(7) There are good relationships between teachers and students.
q44R	(8) There is mutual respect between teachers and students.
Student's Behavior	
Self-Discipline	
q31R	(9) I understand well my teacher's scholastic requirements.
q32R	(10) I know what behavior is allowed or forbidden in school.
q38	(11) This year I was involved in many fights.
q40	(12) Sometimes the teachers treat me badly.
q45R	(13) When I have a problem at school there is always someone I can turn to (from the teaching staff).
Study Efforts	
q47	(14) Weekly hours spent on homework in Math
q48	(15) Weekly hours spent on homework in Hebrew
q49	(16) Weekly hours spent on homework in English
q50	(17) Weekly hours spent on homework in Science and Technology

7 IRTPRO Outputs

Figure 2: Summary Statistics

Item and (Weighted) Summed-Score Statistics for Group 1
Coefficient alpha: 0.6809
Complete data N: 23267

The following Statistics are Computed only for the Listwise-Complete Data:

	Response		With Item Deleted	
			Item-Total	Coefficient
Item	Average	Std. Dev.	Correlation	α
1	3.814	1.273	0.2954	0.6653
2	2.660	1.509	0.3844	0.6532
3	1.002	1.526	0.3116	0.6628
4	0.771	1.206	0.3087	0.6642
5	1.442	1.268	0.3424	0.6602
6	2.906	1.531	0.3628	0.6560
7	1.423	1.259	0.4048	0.6535
8	1.432	1.270	0.3872	0.6553
9	0.939	1.001	0.2300	0.6725
10	0.208	0.599	0.1777	0.6777
11	0.927	1.402	0.1994	0.6763
12	1.730	1.704	0.4082	0.6486
13	0.945	1.349	0.3147	0.6628
14	3.247	1.529	0.1711	0.6809
15	2.516	1.483	0.1290	0.6855
16	3.066	1.604	0.1561	0.6837
17	2.536	1.555	0.1413	0.6849

Figure 3: Goodness of Fit

(a) Unidimensional

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on the loglikelihood	
-2loglikelihood:	1279520.49
Akaike Information Criterion (AIC):	1279724.49
Bayesian Information Criterion (BIC):	1280562.31

(b) Bifactor

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on the loglikelihood	
-2loglikelihood:	1237520.26
Akaike Information Criterion (AIC):	1237758.26
Bayesian Information Criterion (BIC):	1238735.72

Figure 4: Factor Loadings

(a) Unidimensional

Factor Loadings for Group 1

Item	Label	λ_1
1	q34	0.38
2	q37	0.43
3	q39	0.29
4	q41R	0.50
5	q42R	0.60
6	q35	0.43
7	q43R	0.77
8	q44R	0.76
9	q31R	0.49
10	q32R	0.47
11	q38	0.29
12	q40	0.52
13	q45R	0.62
14	q47	-0.27
15	q48	-0.31
16	q49	-0.25
17	q50	-0.28

(b) Bifactor

Factor Loadings for Group 1

Item	Label	λ_1	λ_2	λ_3
1	q34	0.37	0.57	0.00
2	q37	0.42	0.53	0.00
3	q39	0.29	0.25	0.00
4	q41R	0.49	0.00	0.00
5	q42R	0.60	-0.02	0.00
6	q35	0.43	0.60	0.00
7	q43R	0.82	-0.16	0.00
8	q44R	0.79	-0.16	0.00
9	q31R	0.45	0.00	-0.13
10	q32R	0.44	0.00	-0.10
11	q38	0.26	0.00	-0.06
12	q40	0.53	0.00	0.06
13	q45R	0.61	0.00	-0.04
14	q47	-0.15	0.00	0.82
15	q48	-0.20	0.00	0.75
16	q49	-0.14	0.00	0.72
17	q50	-0.18	0.00	0.71

Figure 5: Factor Loadings of Bifactor Model with Subgroups

(a) Low Proportion

Factor Loadings for Group 1

Item	Label	λ_1	λ_2	λ_3
1	q34	0.32	0.57	0.00
2	q37	0.39	0.57	0.00
3	q39	0.30	0.25	0.00
4	q41R	0.48	0.06	0.00
5	q42R	0.61	0.02	0.00
6	q35	0.41	0.59	0.00
7	q43R	0.82	-0.15	0.00
8	q44R	0.79	-0.13	0.00
9	q31R	0.43	0.00	-0.13
10	q32R	0.43	0.00	-0.12
11	q38	0.24	0.00	-0.05
12	q40	0.50	0.00	0.06
13	q45R	0.61	0.00	-0.05
14	q47	-0.16	0.00	0.83
15	q48	-0.18	0.00	0.75
16	q49	-0.14	0.00	0.70
17	q50	-0.18	0.00	0.71

(b) Moderate Proportion

Factor Loadings for Group 2

Item	Label	λ_1	λ_2	λ_3
1	q34	0.39	0.58	0.00
2	q37	0.44	0.52	0.00
3	q39	0.30	0.23	0.00
4	q41R	0.48	-0.04	0.00
5	q42R	0.59	-0.06	0.00
6	q35	0.45	0.62	0.00
7	q43R	0.82	-0.15	0.00
8	q44R	0.80	-0.17	0.00
9	q31R	0.48	0.00	-0.12
10	q32R	0.45	0.00	-0.08
11	q38	0.28	0.00	-0.04
12	q40	0.55	0.00	0.07
13	q45R	0.60	0.00	-0.03
14	q47	-0.17	0.00	0.83
15	q48	-0.22	0.00	0.74
16	q49	-0.16	0.00	0.71
17	q50	-0.18	0.00	0.70

(c) High Proportion

Factor Loadings for Group 3

Item	Label	λ_1	λ_2	λ_3
1	q34	0.37	0.54	0.00
2	q37	0.40	0.51	0.00
3	q39	0.27	0.30	0.00
4	q41R	0.50	0.02	0.00
5	q42R	0.58	-0.01	0.00
6	q35	0.42	0.56	0.00
7	q43R	0.81	-0.21	0.00
8	q44R	0.79	-0.19	0.00
9	q31R	0.43	0.00	-0.15
10	q32R	0.44	0.00	-0.14
11	q38	0.23	0.00	-0.09
12	q40	0.52	0.00	0.05
13	q45R	0.62	0.00	-0.02
14	q47	-0.12	0.00	0.81
15	q48	-0.17	0.00	0.77
16	q49	-0.10	0.00	0.73
17	q50	-0.16	0.00	0.72

Figure 6: Item Parameters - Unidimensional Model

Graded Model Item Parameter Estimates for Group 1, logit: *a*(*θ* - *b*)

Item	Label		*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*	b_4	*s.e.*	b_5	*s.e.*
1	q34	6	0.70	0.01	-5.45	0.12	-4.25	0.09	-2.70	0.06	-1.13	0.03	0.76	0.03
2	q37	12	0.82	0.01	-3.09	0.06	-1.64	0.03	-0.15	0.02	1.07	0.03	2.37	0.04
3	q39	18	0.52	0.02	0.75	0.03	2.05	0.06	3.00	0.09	4.16	0.12	5.48	0.16
4	q41R	24	0.99	0.02	0.48	0.02	1.58	0.03	2.44	0.04	3.42	0.06	4.16	0.07
5	q42R	30	1.27	0.02	-0.97	0.02	0.24	0.01	1.43	0.02	2.55	0.03	3.28	0.05
6	q35	36	0.82	0.02	-3.14	0.06	-1.96	0.04	-0.62	0.02	0.58	0.02	2.04	0.04
7	q43R	42	2.06	0.03	-0.74	0.01	0.23	0.01	1.15	0.01	2.00	0.02	2.51	0.03
8	q44R	48	1.96	0.02	-0.78	0.01	0.24	0.01	1.13	0.01	1.99	0.02	2.53	0.03
9	q31R	54	0.95	0.02	-0.50	0.02	1.28	0.02	2.91	0.05	4.90	0.08	5.55	0.10
10	q32R	60	0.91	0.02	2.15	0.05	3.76	0.08	4.98	0.12	5.99	0.15	6.50	0.18
11	q38	66	0.51	0.02	0.69	0.03	2.11	0.06	3.51	0.10	4.77	0.14	6.30	0.19
12	q40	72	1.03	0.02	-0.75	0.02	0.14	0.01	0.88	0.02	1.62	0.03	2.44	0.04
13	q45R	78	1.35	0.02	0.16	0.01	1.06	0.02	1.74	0.02	2.39	0.03	2.90	0.04
14	q47	84	-0.48	0.02	7.65	0.26	3.26	0.11	1.49	0.06	-0.10	0.03	-1.88	0.07
15	q48	90	-0.55	0.02	5.19	0.15	1.47	0.05	-0.16	0.03	-1.72	0.06	-3.72	0.11
16	q49	96	-0.44	0.02	6.70	0.23	2.94	0.10	1.14	0.05	-0.44	0.04	-2.45	0.09
17	q50	102	-0.50	0.02	4.77	0.14	1.58	0.05	-0.15	0.03	-1.72	0.06	-3.69	0.11

Figure 7: Item Parameters - Bifactor Model

Graded Model Item Parameter Estimates for Group 1, logit: aθ + *c*

Item	Label		a_1	*s.e.*		a_2	*s.e.*		a_3	*s.e.*
1	q34	6	0.85	0.02	7	1.33	0.02		0.00	-----
2	q37	13	0.96	0.02	14	1.21	0.02		0.00	-----
3	q39	20	0.54	0.02	21	0.46	0.02		0.00	-----
4	q41R	27	0.95	0.02	28	0.00	0.02		0.00	-----
5	q42R	34	1.26	0.02	35	-0.05	0.02		0.00	-----
6	q35	41	1.09	0.02	42	1.52	0.03		0.00	-----
7	q43R	48	2.51	0.03	49	-0.50	0.03		0.00	-----
8	q44R	55	2.31	0.03	56	-0.46	0.02		0.00	-----
9	q31R	62	0.88	0.01		0.00	-----	63	-0.26	0.01
10	q32R	69	0.84	0.02		0.00	-----	70	-0.20	0.02
11	q38	76	0.46	0.01		0.00	-----	77	-0.10	0.01
12	q40	83	1.06	0.02		0.00	-----	84	0.12	0.01
13	q45R	90	1.31	0.02		0.00	-----	91	-0.08	0.01
14	q47	97	-0.48	0.02		0.00	-----	98	2.55	0.03
15	q48	104	-0.53	0.02		0.00	-----	105	2.01	0.02
16	q49	111	-0.35	0.02		0.00	-----	112	1.78	0.02
17	q50	118	-0.44	0.02		0.00	-----	119	1.76	0.02

A Additional Data Summaries

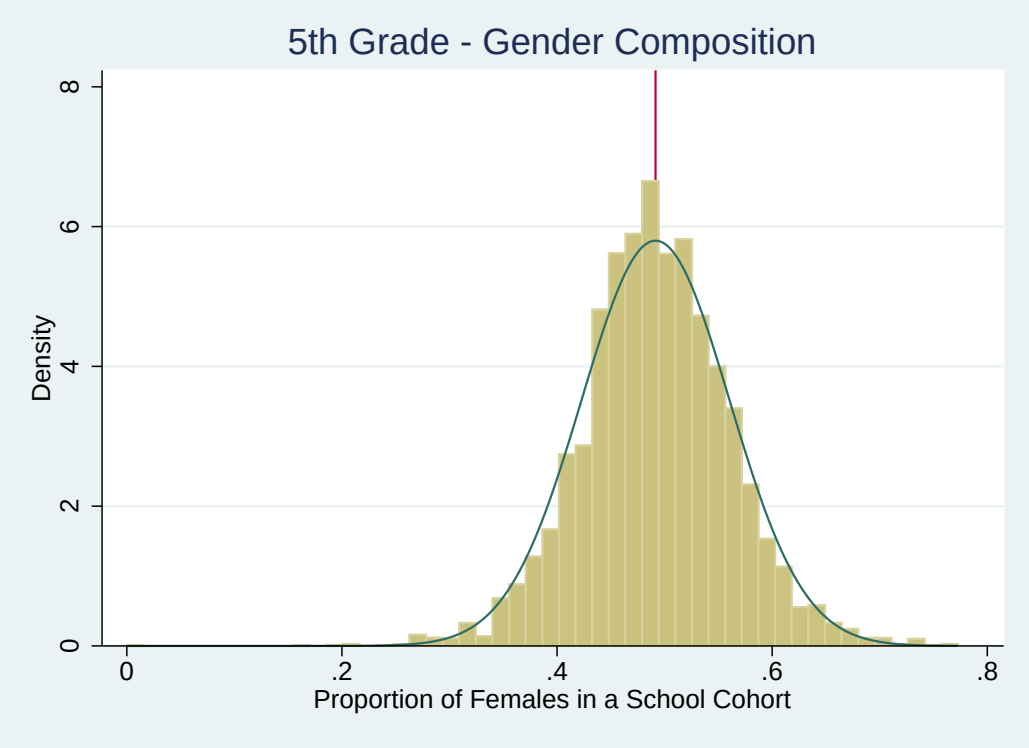
I include additional data summary outputs from Stata as well as IRTPRO for reference.

Figure 8: Summary of the Proportion of Females in a School Cohort for 5th Grade (2002-2005)

prop. female in school cohort

	Percentiles	Smallest		
1%	.3181818	0		
5%	.3793103	0		
10%	.4074074	0	Obs	106119
25%	.4489796	0	Sum of Wgt.	106119
50%	.4893617		Mean	.4915567
		Largest	Std. Dev.	.0687924
75%	.5340909	.7727273		
90%	.5764706	.7727273	Variance	.0047324
95%	.6041667	.7727273	Skewness	-.0400187
99%	.6603774	.7727273	Kurtosis	4.015391

Figure 9: Histogram of the Proportion of Females in a School Cohort for 5th Grade (2002-2005)



B Traditional Statistics

Table of Contents

Item and (Weighted) Summed-Score Statistics for Group 1 **(Back to TOC)**
Coefficient alpha: 0.6809
Complete data N: 23267

The following Statistics are Computed only for the Listwise-Complete Data:

Item	Response Average	Response Std. Dev.	With Item Deleted Item-Total Correlation	With Item Deleted Coefficient α
1	3.814	1.273	0.2954	0.6653
2	2.660	1.509	0.3844	0.6532
3	1.002	1.526	0.3116	0.6628
4	0.771	1.206	0.3087	0.6642
5	1.442	1.268	0.3424	0.6602
6	2.906	1.531	0.3628	0.6560
7	1.423	1.259	0.4048	0.6535
8	1.432	1.270	0.3872	0.6553
9	0.939	1.001	0.2300	0.6725
10	0.208	0.599	0.1777	0.6777
11	0.927	1.402	0.1994	0.6763
12	1.730	1.704	0.4082	0.6486
13	0.945	1.349	0.3147	0.6628
14	3.247	1.529	0.1711	0.6809
15	2.516	1.483	0.1290	0.6855
16	3.066	1.604	0.1561	0.6837
17	2.536	1.555	0.1413	0.6849

Item 1	q34 Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	741	852	2421	4725	7740	10205	597
For listwise-complete data:								
	Frequencies:	652	749	2090	4144	6774	8858	
	Average (wtd) Score:	22.94	23.17	25.99	28.45	31.14	36.01	
	Std. Dev. (wtd) Score:	10.78	8.66	8.15	7.92	7.92	9.27	

Item 2	q37 Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	2464	3646	6314	5707	4457	3986	707
For listwise-complete data:								
	Frequencies:	2182	3236	5567	5013	3857	3412	
	Average (wtd) Score:	23.43	26.09	29.37	32.49	35.44	39.79	
	Std. Dev. (wtd) Score:	8.41	7.80	7.57	7.89	7.98	9.55	

Item 3	q39 Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	15805	3791	2173	1916	1373	1598	625
For listwise-complete data:								
	Frequencies:	13930	3319	1891	1639	1145	1343	
	Average (wtd) Score:	28.75	31.52	34.50	36.98	39.91	43.02	
	Std. Dev. (wtd) Score:	8.65	7.53	8.01	8.26	8.53	9.57	

Item 4	q41R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	15965	5131	2606	1623	619	651	686
For listwise-complete data:								
	Frequencies:	14062	4498	2260	1403	511	533	
	Average (wtd) Score:	28.90	32.69	35.77	39.03	41.63	45.15	
	Std. Dev. (wtd) Score:	8.64	8.26	8.31	8.60	9.63	10.92	

Item 5	q42R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	7420	7526	6573	3303	942	777	740
For listwise-complete data:								
	Frequencies:	6451	6629	5781	2919	821	666	
	Average (wtd) Score:	27.13	29.75	33.07	36.59	40.69	46.24	
	Std. Dev. (wtd) Score:	8.68	8.12	8.18	8.44	8.69	10.53	

Item 6	q35 Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	2380	2690	5128	5757	5781	4893	652
For listwise-complete data:								
	Frequencies:	2098	2385	4535	5051	4993	4205	
	Average (wtd) Score:	22.91	25.56	28.83	31.79	34.39	38.60	
	Std. Dev. (wtd) Score:	8.51	7.60	7.66	7.85	8.27	9.34	

Item 7	q43R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	7441	7758	6509	3176	869	778	750
For listwise-complete data:								
	Frequencies:	6485	6793	5777	2771	764	677	
	Average (wtd) Score:	26.24	29.93	33.50	37.53	41.54	46.75	
	Std. Dev. (wtd) Score:	8.25	7.85	7.96	8.34	8.63	10.19	

Item 8	q44R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	7303	7962	6132	3289	940	820	835
For listwise-complete data:								
	Frequencies:	6403	7023	5424	2913	807	697	
	Average (wtd) Score:	26.40	29.94	33.54	36.83	41.72	46.21	
	Std. Dev. (wtd) Score:	8.46	7.91	7.90	8.25	8.82	10.67	

Item 9	q31R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	10874	9099	4800	1766	172	213	357
For listwise-complete data:								
	Frequencies:	9538	7853	4113	1463	127	173	
	Average (wtd) Score:	28.92	31.34	34.74	38.19	42.90	47.12	
	Std. Dev. (wtd) Score:	9.21	8.42	8.90	9.51	12.64	11.99	

Item 10	q32R Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	22815	2918	761	241	62	107	377
For listwise-complete data:								
	Frequencies:	19875	2451	635	178	49	79	
	Average (wtd) Score:	30.71	35.05	38.94	42.74	43.92	44.85	
	Std. Dev. (wtd) Score:	9.21	9.20	9.00	9.30	10.28	13.31	

Item 11	q38 Category:	0	1	2	3	4	5	(Back) Missing
	Frequencies:	15530	4042	2967	1738	1209	1158	637
For listwise-complete data:								
	Frequencies:	13828	3521	2538	1478	980	922	
	Average (wtd) Score:	29.34	32.03	34.42	36.32	38.23	40.59	
	Std. Dev. (wtd) Score:	9.04	8.50	8.72	8.91	8.27	10.17	

Item	q40							(Back)
12	Category:	0	1	2	3	4	5	Missing
	Frequencies:	9172	4896	3989	3329	2493	2782	620
For listwise-complete data:								
	Frequencies:	8097	4300	3493	2901	2123	2353	
	Average (wtd) Score:	26.32	29.36	32.22	35.28	38.05	42.22	
	Std. Dev. (wtd) Score:	7.87	7.35	7.54	7.84	8.14	9.37	

Item	q45R							(Back)
13	Category:	0	1	2	3	4	5	Missing
	Frequencies:	14483	5559	2978	1773	816	1073	599
For listwise-complete data:								
	Frequencies:	12723	4808	2565	1557	700	914	
	Average (wtd) Score:	28.52	31.98	35.21	38.19	39.68	43.99	
	Std. Dev. (wtd) Score:	8.48	8.16	8.47	8.51	9.10	10.56	

Item	q47							(Back)
14	Category:	0	1	2	3	4	5	Missing
	Frequencies:	722	4085	4070	4638	5100	7926	740
For listwise-complete data:								
	Frequencies:	634	3568	3567	4082	4477	6939	
	Average (wtd) Score:	27.43	26.81	28.72	30.89	33.15	35.22	
	Std. Dev. (wtd) Score:	10.55	9.09	8.77	8.90	8.49	9.20	

Item	q48							(Back)
15	Category:	0	1	2	3	4	5	Missing
	Frequencies:	1581	6723	5311	4964	4357	3304	1041
For listwise-complete data:								
	Frequencies:	1394	5982	4748	4396	3856	2891	
	Average (wtd) Score:	28.44	28.42	30.39	32.75	34.41	35.88	
	Std. Dev. (wtd) Score:	10.20	9.21	8.89	8.86	8.67	9.61	

Item	q49							(Back)
16	Category:	0	1	2	3	4	5	Missing
	Frequencies:	1436	4468	4173	4324	5084	7012	784
For listwise-complete data:								
	Frequencies:	1256	3943	3651	3776	4449	6192	
	Average (wtd) Score:	28.05	27.31	28.91	31.33	33.18	35.54	
	Std. Dev. (wtd) Score:	9.99	8.81	8.94	8.78	8.69	9.23	

Item	q50							(Back)
17	Category:	0	1	2	3	4	5	Missing
	Frequencies:	2367	6019	5254	4673	4257	3852	859
For listwise-complete data:								
	Frequencies:	2005	5317	4665	4157	3739	3384	
	Average (wtd) Score:	28.60	28.07	30.28	32.61	34.30	36.27	
	Std. Dev. (wtd) Score:	9.84	9.08	9.00	8.53	8.72	9.51	

Summary of the Data and Control Parameters **(Back to TOC)**

Sample Size	27281
Number of Items	17