# An Analysis of Gender-Based Differential Item Functioning in the PISA 2018 and 2022 Cycles

Isabel Xiong
*Washtenaw International High School*
Ann Arbor, MI United States
shuey.rong@gmail.com

***Abstract*—The Programme for International Student Assessment (PISA), a large-scale international assessment, has become increasingly important in informing policy decisions by the Organization for Economic Co-operation and Development (OECD) and national education agencies. The integrity of the findings derived from its data is contingent upon the fairness and validity of its test items. This study investigates gender-based differential item functioning (DIF), an indicator of potential test bias, in the mathematics questions administered to test-takers in the United States in the 2018 and 2022 PISA cycles. A multi-group Item Response Theory (IRT) framework was employed with the Wald and Likelihood Ratio Test. A multi-trial iterative anchor selection procedure was implemented to ensure higher reliability of the test results. The analysis reveals that the 2022 PISA mathematics test exhibits a higher rate of items with potential DIF compared to the 2018 cycle, suggesting test bias is not simply a result of outdated questions or screening methods. Further investigation of the items flagged in both cycles suggests that gender-based bias does not strictly favor one gender, changing depending on the item and potentially the ability level of the test-taker, but leans slightly towards male test-takers. These findings underscore the persistent challenge of gender-based bias in standardized assessments, but also highlight how it can be detected and mitigated.**

***Index Terms*—Differential Item Functioning, Item Response Theory, Gender-based bias**

## I. Introduction

Standardized tests have been widely used in modern education systems for evaluating student performance. In turn, statistical analysis of the test results provide critical insight for informing policies and resource allocation in education. The Programme for International Student Assessment (PISA), administered by the Organization for Economic Co-operation and Development (OECD), is one of the most influential tests, with 81 participating countries as of 2022. It provides a measure of the mathematics, reading, and science skills of 15-year olds around the world. Demographic information is collected, including gender, parental education levels, and socioeconomic status, which allows the study of correlations between academic performance and a range of background factors [1]. Results from these studies have been increasingly influential on educational policies and resource allocation [2], [3]. It is therefore crucial that the test accurately reflects a student's abilities and is free from systematic bias [4], [5]. In this study, we investigate and compare potential gender bias in the mathematics test items in the 2018 and 2022 PISA cycles, using a multi-group Item Response Theory (IRT) framework, combined with a rigorous iterative anchor selection process, to provide high reliability in the analysis.

## II. Methodology

### A. Item Response Theory (IRT) Framework

Here, Item Response Theory (IRT) models are employed to investigate differential item functioning. These models are designed to explain the relation between an individual's level of a latent trait (represented by $\theta$), such as mathematical ability, and their responses to test items, which can be directly measured [6]. The relationship between an item and the latent trait is modeled by the Item Characteristic Curve (ICC), which plots the probability of a correct response as a function of the latent trait level, standardized to a scale ranging from -4 to +4 [1]. An item's ICC is determined by a set of item parameters: discrimination ($a$), reflecting the item's ability to differentiate between test-takers with different levels of ability [7]; difficulty ($b$), representing the specific ability level ($\theta$) at which a test-taker has a 50% probability of answering the item correctly; and the guessing parameter ($c$), which reflects the probability of correctly guessing the answer [7] and determines the lower asymptote of the ICC, where test-takers with lower levels of the latent trait may answer correctly by chance. The $c$ parameter was not considered in this study, in order to improve computational efficiency and because of its limited impact on the ICC.

### B. Differential Item Functioning (DIF) Detection

Item bias was detected by looking for Differential Item Functioning (DIF). DIF occurs when test-takers from different groups (e.g., male and female) with the same level of the latent trait have a different probability of answering an item correctly.

An item's level of DIF can be tested for using IRT-based logistic regression models, which assess whether group membership adds explanatory power to the probability of a correct response beyond the test-taker's ability level ($\theta$) [7]. Two primary statistical tests are commonly employed for this purpose: the Wald Test and Likelihood Ratio Test (LRT).

The Wald Test fits a single, unconstrained model that assumes group membership has an impact on test-taker's responses [8]. It then assesses how significant the effect of group membership is on item performance, with a higher Wald statistic indicating a more significant impact.

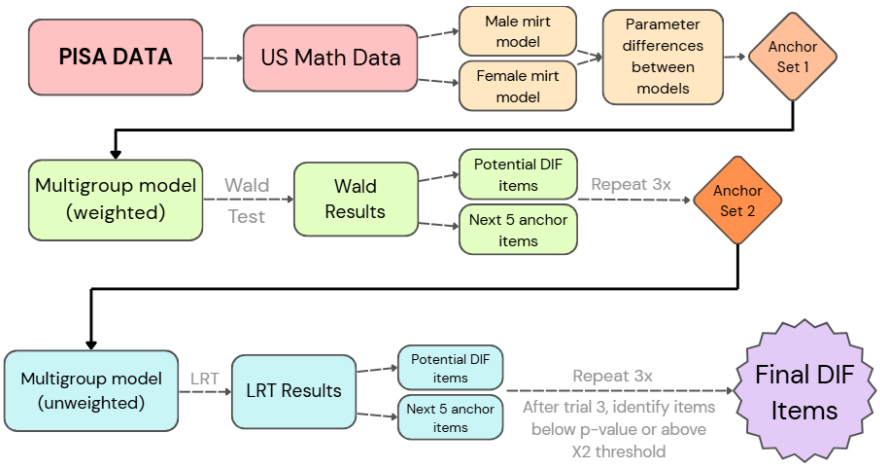


Fig. 1. An overview of the data analysis procedure used in this investigation.

The LRT compares the goodness-of-fit of two nested models. The first, a more constrained model, assumes no DIF by forcing the item parameters to be the same across groups. The second, a more complex model, allows the parameters to vary between groups. If the difference in goodness-of-fit to the data between the two models is statistically significant (based on a chi-square test), the more complex model with varying parameters is considered a better fit, and the null hypothesis of no DIF can be rejected [8]. The LRT is considered more powerful and reliable than the Wald test, but is more computationally demanding. In this investigation, the Wald test was used for initial screening and anchor identification, and the LRT for testing flagged items and actually identifying DIF.

## III. Data Processing and Analysis

In order to create the dataset for analysis, PISA data was collected and cleaned, with math item-responses from specifically US respondents extracted. Then, an iterative anchor selection process was employed: an initial Anchor Set 1 was created by analyzing differences in item parameters for male and female models, and used to create a comprehensive and accurate Multigroup model. The Wald Test was run on this model, with multiple trials, to identify Anchor Set 2. Finally, Anchor Set 2 was used to create an new Multigroup model, and three trials were performed with the Likelihood Ratio Test to determine the final set of DIF items. All data processing and analysis was done in Rstudio, using the mirt package for the analysis. The full methodology is summarized in Fig. 1.

### A. The Programme for International Student Assessment

The Programme for International Student Assessment (PISA) is an international assessment of the mathematics, reading, and science skills of 15-year-olds. It is administered every three years, with the exception of the cycle interrupted by the global pandemic. The PISA 2018 and PISA 2022 cycles are the most recent to be completed. In PISA 2022, 81 countries and economies participated [1]. The PISA 2022 mathematics assessment featured a hybrid multistage adaptive test design (MSAT), tailoring items to a student's performance level [1]. The adaptive nature of the test design means a robust psychometric framework like IRT is necessary to ensure accurate and unbiased ability estimates.

### B. Data Collection and Cleaning

This is summarized in the red sections of Fig 1. The data used for this study was obtained from the official PISA website, a publicly accessible resource. The relevant files, in SAS format, were the PISA student questionnaire and cognitive item data files for both the 2018 and 2022 cycles. The specific variables used were item responses, final student weights, gender, country ID, and student ID. For this investigation, only questions from the digital version of PISA were used, to control for differences in test administration between paper and digital versions, which could also impact item functioning. The data were imported into RStudio using the Haven package.

Only data from US respondents were used to control for country-specific cultural and educational factors that might otherwise act as confounding variables to the analysis. Non-dichotomous items were removed from the dataset, since the DIF detection methods were specifically designed for binary (correct/incorrect) response data, as were participants with no responses to the designated set of items. This was done to ensure they wouldn't skew the results of the Likelihood Ratio Test, which can be susceptible to estimation errors when there are limited responses for a certain participant or item [9]. Finally, the survey weights were normalized to sum to the sample size.

### C. Iterative Anchor Selection

All DIF analyses were conducted using the mirt package. The multipleGroup function within mirt, which is specifically designed to perform multi-group IRT analyses, was used to create and fit the models necessary for the DIF tests.

An essential part of DIF analysis is a set of anchor items assumed to be free of DIF [10]. They serve to align the latent ability scales of the groups being compared (in this case, males

and females). Without this alignment, the item parameters from the different groups would not be on a common scale, making a direct comparison impossible; this could lead to both false positives and DIF not being detected at all. Thus, a contaminated anchor set, including items that actually have DIF, can greatly skew the results of either the LRT or Wald. This risk is heightened with larger anchor sets, but a too-short anchor can decrease the statistical power of the DIF tests. An anchor length of three to five items has been suggested as an appropriate balance [10]. Five anchors were used in this study to account for the number of items and provide increased DIF detection power.

To address the risk of anchor contamination and ensure the robustness of the DIF findings, an iterative anchor selection procedure was used. This purification process, a recognized strategy in psychometrics [10], involves multiple trials to converge on a stable set of DIF-free items. These repeated trials can also remove any items that may have been incorrectly assumed to be DIF-free in the initial anchor set.

The procedure for the set of items from each PISA cycle involves two stages, described below:

- First, separate mirt models were created for the male and female test-takers. The absolute value of the differences between the estimated $a$ and $b$ parameters for each item were then calculated and summed. The five items with the smallest differences in parameters were selected as the initial anchor set [10], labeled Anchor Set 1. The initial anchors identified for the PISA 2018 cycle were items number 6, 30, 32, 40, and 48; for PISA 2022, items number 6, 31, 38, 40, and 49 were used. Fig 2. provides a visualization of the range of these parameter differences, and the orange sections of Fig 1. provides a visualization of this process.
- Second, a multipleGroup model was fit using these initial anchors, and the Wald test was run to test for DIF on all other items. Any potential DIF items above the X2 critical value, or below the p-value threshold, were noted down; the five items with the highest p-values from this initial test were used as anchors for the next multipleGroup model. This process was repeated twice, for a total of three trials. Based on the results of the final trial, the five items with the highest p-value were designated Anchor Set 2. This process is shown in the green section of Fig 1.

### D. DIF Detection Criteria

After identification of a robust anchor set, the Likelihood Ratio Test was used to determine which items had DIF. Similar to initial anchor identification, this was done over three trials. An initial multipleGroup model, unweighted, was created using Anchor Set 2. Weights were not included for this model because differences in model fit detected by the LRT were greatly exaggerated when responses were weighted differently. The LRT was run on this unweighted model, and the 5 items with the highest p-value were then used to create the second generation model. Three trials were run, and the results of the third LRT were used to determine DIF items.

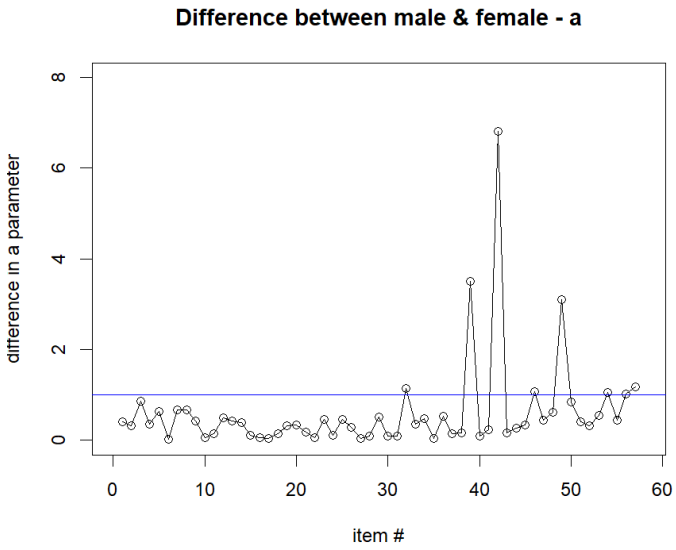


Fig. 2. Difference in a-parameter between male and female groups, 2022 PISA cycle. Line at y=1 is added for scale. Items 6, 31, 38, 40, and 49 were used as anchors, after considering both the $a$ and $b$ parameter.

The main measure used to identify DIF was the p-value, adjusted from $0.05\%$ using the Bonferroni correction. The p-value threshold was approximately $0.098\%$ for 2018 data, and approximately $0.088\%$ for 2022. To confirm these findings, the chi-square value, adjusted to 13.855 for 2018 and 14.078 for 2022 to match the adjusted p-values, was also considered. In the conclusions, both the adjusted and original critical values were considered.

## IV. Results

### A. Item Parameter Differences

A preliminary analysis of the PISA 2018 and 2022 math items provides an initial visualization of potential item-level differences between male and female test-takers. The absolute value of the difference between both the $a$ and $b$-parameters for male and female groups was calculated for each item; this helped determine which items were estimated to have similar parameters between male and female test-takers, and thus which items could serve as initial anchors. As an example, the graph for differences in the a-parameter for 2022 items is shown in Fig. 2. Each point represents the difference in the parameter for an individual item. Items 6, 9, 31, 38, and 40 were used as Anchor Set 1. For the 2018 cycle, Anchor Set 1 consisted of items number 6, 30, 32, 40, and 48.

### B. Anchor Selection Findings

For selection of Anchor Set 2, the results of the Wald tests run on weighted multipleGroup models are summarized in Tables I and II below for the 2018 and 2022 cycles respectively. The items with the highest p-value in the third trial were used as initial anchors for the unweighted model tested with the Likelihood Ratio Test.

TABLE I
2018 ANCHOR IDENTIFICATION

| | ***Trial 1*** | ***Trial 2*** | ***Trial 3*** |
|---|---|---|---|
| W $>$ 5.991 | 18, 7 | 18, 7 | 7 |
| p $<$ 0.098% | None | None | None |
| Highest p | 26, 9, 27, 31, 19 | 32, 30, 48, 40, 2 | 9, 31, 27, 45, 26* |

∗These 5 items were identified as Anchor Set 2

TABLE II
2022 ANCHOR IDENTIFICATION

| | ***Trial 1*** | ***Trial 2*** | ***Trial 3*** |
|---|---|---|---|
| W $>$ 5.991 | 35, 46, 14, 12, 13, 21, 23 | 35, 46, 13, 12, 23, 24 | 14, 13, 35, 12, 46, 23, 1, 21 |
| p $<$ 0.088% | None | None | None |
| Highest p | 19, 18, 42, 10, 22 | 4, 53, 49, 7, 47 | 9, 31, 27, 45, 26* |

∗These 5 items were identified as Anchor Set 2

### C. Final DIF Analysis

The results of the DIF analyses for the 2018 and 2022 cycles are summarized in Tables III and IV. The table presents the items identified as potentially having DIF over three trials using different anchors, determined by whether they reach the critical value for the X2 statistic, or are below the threshold for the p-value. The initial anchors used were those identified as Anchor Set 2.

TABLE III
PISA 2018 MATH ITEMS WITH DIFFERENTIAL ITEM FUNCTIONING

| | ***Trial 1*** | ***Trial 2*** | ***Trial 3*** |
|---|---|---|---|
| X2 $>$ 5.991* | 7 | 7,33 | 7,33 |
| p-value $<$ 0.098% | None | None | None |
| Highest p-value | 5, 32, 50, 40, 51 | 27, 9, 36, 2, 31 | 5, 50, 50, 32, 51 |

*No items were over the 13.855 threshold.

TABLE IV
PISA 2022 MATH ITEMS WITH DIFFERENTIAL ITEM FUNCTIONING

| | ***Trial 1*** | ***Trial 2*** | ***Trial 3*** |
|---|---|---|---|
| X2 $>$ 5.991 | 46*, 12,13, 35, 1, 23, 14 | 46*, 12, 13, 35, 14, 1, 55, 23, 21 | 46*, 13, 35, 14, 12 |
| p-value $<$ 0.088% | 46 | 46 | 46 |
| Highest p-value | 5, 32, 50, 40, 51 | 27, 9, 36, 2, 31 | 5, 50, 50, 32, 51 |

*Item over the X2 = 14.078 threshold.

### D. Comparative Analysis

A direct comparison between confirmed DIF rates of 2018 and 2022 is shown in Table V. This provides a clear answer to the central research query, demonstrating that 2022 displayed a higher rate of confirmed DIF, with 1.75% in 2022 compared to no confirmed DIF items in 2018. The rate of potential DIF items, those passing the unadjusted X2 or p-value thresholds but not the adjusted ones, was also higher in 2022, at 8.78% compared to 3.92% in 2018.

TABLE V
RATES OF DIFFERENTIAL ITEM FUNCTIONING, 2018 AND 2022

| **Year** | ***Number of DIF items*** | ***Percentage DIF*** | ***% of Potential DIF items*** |
|---|---|---|---|
| 2018 | 0 | 0% | 3.92% |
| 2022 | 1 | 1.75% | 8.78% |

TABLE VI
TYPES OF DIF IN 2018 MATH ITEMS

| Item Number | Type of DIF | Group favored |
|---|---|---|
| 7 | Uniform | Male |
| 33 | Non-uniform | High $\theta$: female |
| | | Low $\theta$: male |

These results show clear difference in the prevalence of gender-based DIF in mathematics items across the two PISA cycles. Uniform DIF tended to favor males. However, this bias doesn't entirely favor one group. By overlaying a given item's ICCs for males and females, it becomes clear that some of the DIF is nonuniform as well, favoring different genders at different skill levels. In Fig 3., the overlaid ICCs show that males are more likely to answer correctly than females at the same low skill levels, while females are more likely to answer correctly when both are at high skill levels, thus demonstrating non-uniform DIF. In the case of uniform DIF, the two ICCs would not intersect, except at the asymptotes. As shown in Tables VI and VII, there was a roughly equal split between uniform and nonuniform DIF in the potential DIF items, and uniform DIF tended to favor males.

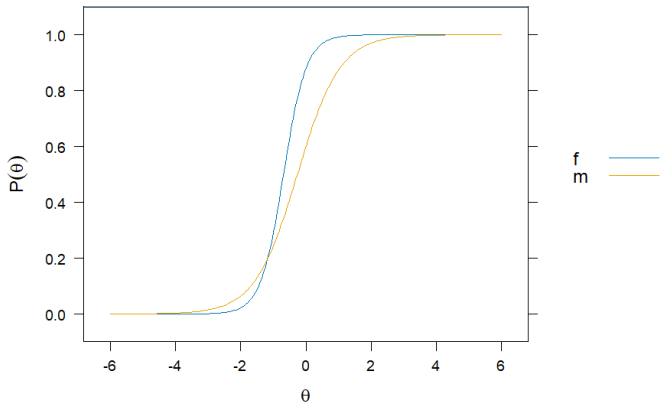


Fig. 3. The overlaying of ICCs, plotted as the probability of a correct response ($P(\theta)$) as a function of the latent trait level ($\theta$), for male (m) and female (f) test-takers for item 46 from the 2022 cycle.

## V. DISCUSSION

### A. Conclusions

Based on the findings of this investigation, the PISA 2022 cycle had a notably higher rate of gender-based DIF in its mathematics items compared to the PISA 2022 cycle. This suggests that outdated methods of items design or test administration are likely not the primary reason for observed

TABLE VII
TYPES OF DIF IN 2022 MATH ITEMS

| Item Number | Type of DIF | Group favored |
|---|---|---|
| 35 | Uniform | Female |
| 14 | Uniform | Male |
| 12 | Uniform | Male |
| 46 | Non-uniform | High $\theta$: female |
| | | Low $\theta$: male |
| 13 | Non-uniform | High $\theta$: male |
| | | Low $\theta$: female |

DIF. The presence of DIF suggests that observed differences in mean scores between male and female students are not solely due to a true ability gap but could be, at least in part, caused by item bias. The mix of uniform and non-uniform DIF, as well as the favoring of males, suggests random issues with item design as well as consistent or systemic bias. This raises concerns about the usage of PISA scores to make high-stakes policy decisions, especially without fully accounting for these item-level inequalities.

The findings contribute to a broader understanding of test fairness in large-scale standardized tests. The increase in the rate of confirmed and potential DIF items between 2018 and 2022 suggests that an increase in the number of items, a result of math being the focus of the 2022 test cycle, introduced more items with DIF. However, it is important to note that many of the items identified as having or potentially having DIF, including items 7, 33, and 46, were present in both cycles. This suggests that DIF did not come solely from the items themselves, but that a changing educational and cultural landscape could have made minor imbalances in certain items more extreme. Especially with the two cycles being before and after COVID-19, a multitude of factors, including different formats for virtual schooling, existed to increase the variance in skill distribution even within a single country. This investigation highlights further the importance of a nuanced approach in properly assessing DIF.

### *B. Further Extensions*

The findings of this report have limitations. A persistent challenge is the difficulty in identifying a set of truly DIF-free anchor items [11]. While the iterative purification procedure employed here mitigates the risk of anchor contamination, there is a possibility that the final set of anchors included a subtle DIF item, potentially skewing the results for other items. More precise and reliable methods exist that would offer a higher degree of certainty and could potentially be used in future investigation: methods using the multiple causes multiple indicators (MIMIC) model for DIF have been shown to accurately estimate the DIF effects of individual items without prior knowledge of an anchor set, and allow for Type-I error control [12]. Regularization methods such as the lasso EMM can also be used for an effective analysis without the need for an iterative anchor selection procedure [13], provided the sample size is sufficiently large, and used on multidimensional IRT models [13], which measure multiple latent traits simultaneously. With this method, PISA's measurement of reading, science and mathematics skills could potentially be analyzed, with less risk of anchor contamination, and with a single analysis.

Secondly, in this investigation, DIF was only identified quantitatively, using data analysis. A more qualitative examination, such as an analysis of the contents of the specific DIF-identified items, would help explain not only which items display DIF, but the reasoning behind it. Previous research has documented that male students often display a clear advantage on "Space and Shape" items and on "complex multiple-choice items" in mathematics [14]. A future study and consultation with educational experts could show whether the DIF-flagged items in this investigation are overrepresented in certain domains. Such an analysis would help pinpoint the source of the gender bias and provide ways for test developers to revise existing items and develop new ones in a fairer way.

## ACKNOWLEDGMENT

I would like to thank Professor Gongjun Xu at the University of Michigan, Ann Arbor for his guidance and support during my investigation.

## REFERENCES

[1] OECD. Pisa 2022 technical report. Technical report, Paris, 2024.
[2] Sotiria Grek. Governing by numbers: the pisa 'effect' in europe. *Journal of Education Policy*, 24(1):23–37, 2009.
[3] S. Sellar and B. Lingard. The oecd and the expansion of pisa: new global modes of governance in education. *British Educational Research Journal*, 40(6):917–936, 2014.
[4] Laura C. Engel, David Rutkowski, and Greg Thompson. Toward an international measure of global competence? a critical look at the pisa 2018 framework. *Globalisation, Societies and Education*, 17(2):117–131, 2019.
[5] Daniel Bart. Research discourse in the programme for international student assessment: A critical perspective. *European Educational Research Journal*, 23(1):145–162, 2022.
[6] Frank B. Baker. *The Basics of Item Response Theory. Second Edition*. ERIC Clearinghouse on Assessment and Evaluation, 2001. ISBN: 9781886047037 ERIC Number: ED458219.
[7] Frances M. YANG and Solon T. KAO. Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, 26(3):171–177, June 2014.
[8] Wainer H. Holland, P. W. *Differential item functioning*. Lawrence Erlbaum Associates, Inc., 1993.
[9] Bruno Dujardin, Jef Van den Ende, Alfons Van Gompel, Jean-Pierre Unger, and Patrick Van der Stuyft. Likelihood ratios: A real improvement for clinical decision making? *European Journal of Epidemiology*, 10(1):29–36, February 1994.
[10] Julia Kopf, Achim Zeileis, and Carolin Strobl. Anchor Selection Strategies for DIF Analysis. *Educational and Psychological Measurement*, 75(1):22–56, February 2015.
[11] Gabriel Wallin, Yunxiao Chen, and Irini Moustaki. DIF Analysis with Unknown Groups and Anchor Items. *Psychometrika*, 89(1):267–295, March 2024.
[12] Li C. Ouyang J. Xu G. Chen, Y. Dif statistical inference without knowing anchoring items. *Psychometrika*, 88(4):1097–1122, 2023.
[13] Zhu R. Xu G. Wang, C. Using lasso and adaptive lasso to identify dif in multidimensional 2pl models. *Multivariate behavioral research*, 58(2):387–407, 2023.
[14] O. L. Liu and M. Wilson. Gender differences in large-scale math assessments: Pisa trend 2000 and 2003. *Applied Measurement in Education*, 22(2):164–184, 2009.