A Mixture Model Analysis of DIF on a Test of Motor Development

Allan S. Cohen
University of Georgia
October 19, 2007

Cooper Institute
Diversity in …

Definition of DIF

- DIF is said to arise when the item measures some nuisance variability not part of the intended construct.

- DIF is conditional, that is, it is defined as a differential propensity for a particular response for examinees of the same ability but from different groups.

DIF is typically defined based on manifest group characteristics.

- Most methods associate some manifest group characteristic such as gender or ethnicity with differential item performance.

- These characteristics are usually only weakly related to DIF.

- Further, this approach does not readily contribute to understanding what may be the cause(s) of DIF.

Fundamental Assumption In this Study

- We begin with the assumption that not all examinees in one group are consistently advantaged (or disadvantaged) by a DIF item.
- Instead, we assume that examinees advantaged or disadvantaged by DIF items are better regarded as latent classes in the data.
- We then use mixture IRT models for identifying latent groups of examinees that are advantaged or disadvantaged by DIF in particular items.

Illustrate Two Types of DIF Analyses

- In Study 1, we analyze item response patterns for (a) gender DIF and (b) ethnic group DIF, and then describe a mixture IRT model for each DIF analysis, that illustrates how the manifest characteristic associated with DIF often has a weak relationship with the groups actually advantaged or disadvantaged by the item.
- In Studies 2 & 3, we illustrate an alternative approach to DIF assessment that first defines the dimension by which DIF occurs and then looks at examinee and item characteristics associated with that dimension.

Use of Mixture IRT Models

Mixture IRT models provide a useful means of investigating how qualitative examinee differences (e.g., use of different cognitive strategies or different problem solving strategies) might lead to differences in responses to test items

Mixture Rasch Model

- The Rasch model is assumed to hold for each latent class, but the item parameters differ for different latent classes, and
- each examinee is parameterized by a class membership (g) and a within class ability parameter (θ_g).

$$P(U = 1 | g, \theta_g) = \frac{\exp(\theta_g - b_{ig})}{1 + \exp(\theta_g - b_{ig})}$$

Study 1: Gender DIF & Ethnic Group DIF Analysis on a Test of Motor Development

- In a gender DIF analysis, all we really learn is that one group, conditional on ability, responded differently to an item than did the other group.
- In this study, we use a **two-group mixture Rasch model** (MRM) for identifying latent groups of examinees that are advantaged or disadvantaged by DIF in particular items.
- We show that, even when we detect gender DIF, a large proportion of the sample can be found to respond more like members of the other gender.

Measure

- Test: Test of Gross Motor Development (Ulrich, 1985)
- 45 items on two types of motor behavior tasks:
  - locomotor items (running, galloping, hopping, leaping, jumping, skipping, sliding); and
  - object control items (striking, bouncing, catching, kicking, and throwing).
- Examinees perform three trials of each behavior. Scored 1 if 2 of 3 performances met criterion.

Sample

- Data from norming sample used (N = 909).
- Multistage sample: nine states ; 451 boys and 458 girls; ages three to ten
  - . A minimum of 100 children in each of the eight age groups.
- Ethnic groups in the sample: 143 (15.7 %) African-Americans, 55 (6.1 %) other minorities, 711 (78.2 %) Caucasian.
- Twenty-seven (3.0 %) of the children had some type of disability.

Mantel Haenszel DIF Analysis

- Gender: 19 DIF items (9 favored males, 10 favored females)
  - 8 of 9 male-DIF items (i.e., harder for females) were object control items
  - 10 female-DIF items were locomotor items
- Ethnicity: 2 DIF items (1 favored African-Americans, 1 favored Caucasians)
  - Both items were object control items

Estimating Rasch Item Difficulties

- Pattern of DIF use to set up MULTILOG calibration runs to estimate Rasch item difficulties for the two gender and two ethnic groups.
- Non-DIF items were constrained to be equal in both groups; DIF items were unconstrained in both groups.

MCMC Algorithm to Estimate Mixture Rasch Model

- Rasch item difficulties used to set up two-class MRM.
- Estimation of latent class membership for the two-class MRM was done using a Markov chain Monte Carlo (MCMC) estimation algorithm with item parameters fixed at values estimated from sample based on gender DIF results.
- If the latent classes correspond directly to gender, males would be classed in Class 1 and females in Class 2.

Latent Class Assignment Based on Pattern of Gender DIF

Latent Class	Gender		Total Sample
	Males	Females	
Class 1	276	198	474
Class 2	175	260	435
Total	451	458	909

- 59.0 percent of examinees assigned to latent class associated with their gender.

Latent Class Assignment Based on Pattern of Ethnic Group DIF

Latent Class	Ethnicity: Cauca-sians	African-Americans	Total
Class 1	353	57	410
Class 2	358	86	444
Total	711	143	854

- 51.4 percent of examinees assigned to latent class associated with their ethnic group membership.

Age & Disability Status

- In the gender analysis, males and females did not differ in age or disability status, but Class 1 was about 3 years older than Class 2.

- In the ethnic group analysis, Caucasians and African-Americans did not differ in age or disability status, but Class 1 was about 4 years younger than Class 2.

Conclusions for Study 1

- The pattern of DIF did not accurately identify those individuals for whom the items appear to have functioned differentially.
- The MRM analysis indicated latent classes of examinees, defined based on response propensities, did not correspond well to gender or ethnic groups.

Study 2: Two-Class Mixture Nominal Response Model

- First DIF was identified as a secondary nuisance dimension and then manifest examinee characteristics were associated with that dimension.
  - Fit a mixture nominal response model to responses to 12 English Usage items.
  - Then classified examinees in the sample using a two-class solution and examined characteristics of these examinees
    - to gain insight into the potential differences between manifest groups of these examinees that would help explain why this DIF occurred.

1. *I, Claudius*, one of television's most lauded series, are being rebroadcast.
1 2 3 4
No error. (Answer: 4)
5

2. Maria, who had just eaten, thought concerning having a candy bar or ice
1 2 3 4
cream. No error. (Answer: 3)
5

3. Nobody believes that the defendant will be acquitted, even his strongest
1 2 3
supporters are convinced of his guilt. No error. (Answer: 3)
4 5

Results for Study 2

- Two distinct classes were identified:
  - Punctuation Class (about 72%)
  - Usage Class (about 28%)

- These two classes clearly exhibit different cognitive strategies in answering the usage items.

- Next, we asked whether these two classes are associated with manifest examinee characteristics.

Data for Study 2

- Responses of 5,242 examinees who entered a university campus in 2000/01.
  - 2,297 males
  - 2,933 females
  - 12 failed to indicate gender
- Gender, age, and language spoken in the home are self-report

Achievement-Related Effect Sizes in Study 2

Variable	Usage	Punctuation
English Usage		.31
ACT-Composite		.15
ACT-English		.16
ACT-Math		.11
1st Semester GPA		.13
1st Semester GPA Credits		.08
Cumulative GPA		.13
Total Credits		.14
Transfer Credits		.10
Failure Credits		.09

Summary of Study 2

- No gender or age differences in latent groups
- English was dominant language spoken in the home for both groups.
- Some ethnic group differences may be present but numbers for most groups are small:
  - Usage Group: African-Americans: (about 41%), other groups between 28% and 35%)

Study 3: MRM Analysis of TGMD Data

- In this study, we first identify the secondary dimension that was assumed to be causing the DIF
  - using an exploratory analysis, we fit multiple models to find the one best fitting model.
- We then examined the items to see which ones functioned differentially among the latent classes.

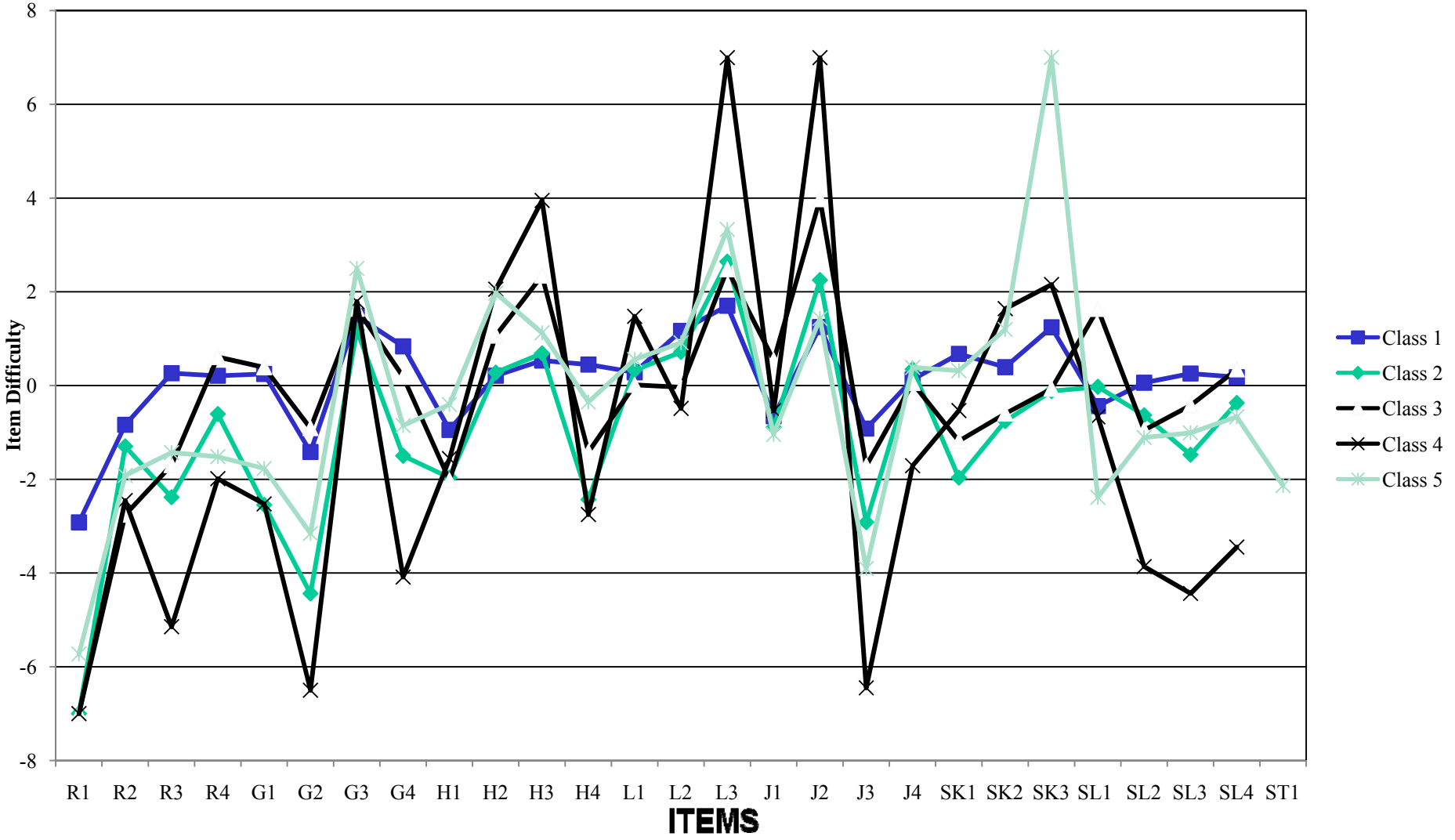
Methods

- Test: Same test used in Study 1, the TGMD.
- Sample: Same norming sample used in Study 1.
- Estimation of MRM: computer program Winmira

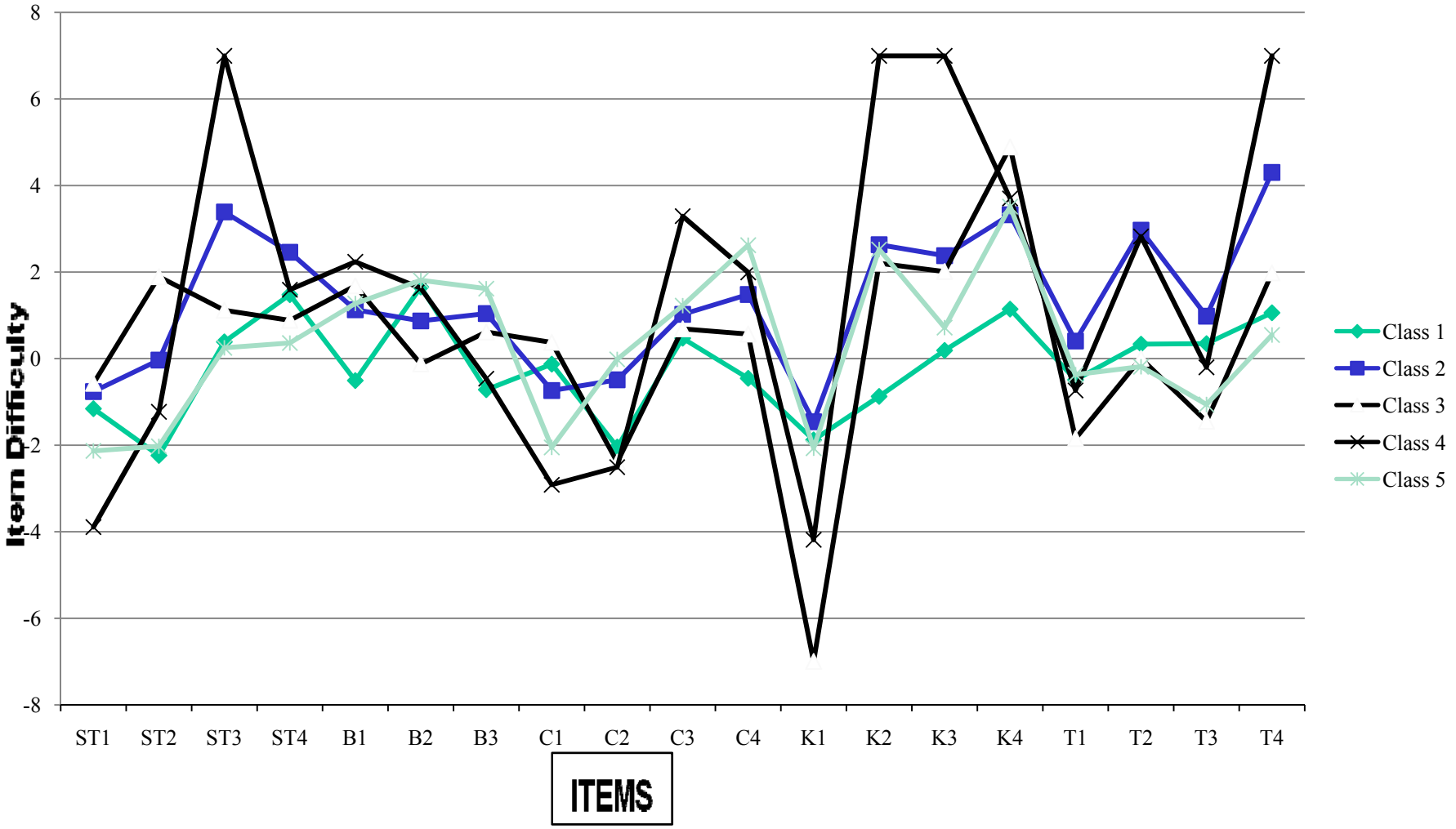
Results for Study 3

- A 5-class model was fit to all 45 items.
- We interpret that model.

Profile of Item Difficulties For Locomotor Items



Profile of Item Difficulties For Object Control Items



Patterns of Differences Among Latent Classes

- The groups appear to differ depending to some extent on positioning required of the arms, the trunk or the legs for performing tasks, for either locomotor or object control.
- The relative difficulty of these tasks for one group compared to another groups helps to amplify possible reasons for differences in performance.

Comparison of Some Propensities

Hard

- 1 Locomotor, positioning
- 2 Weight transfer, arm mechanics
- 3 Arm & trunk positioning, & feet off ground
- 4 Arm & Trunk positioning
- 5 Proper arm, elbow, leg positioning, e.g., while swinging

Easy

- 1 Object control, mechanics (e.g., arm positioning, follow through)
- 2 Locomotor, feet off ground
- 3 Arm movements, weight transfer
- 4 Running, galloping, correct position for jumping, leaping.
- 5 Trunk rotation e.g., for sliding, throwing

Descriptive Analysis of the Five Latent Classes

Latent Class	Raw Score Mean (SD)	Prop.	Age (SD)
1	38.1 (5.3)	.248	8.2 (1.7)
2	29.2 (7.1)	.226	6.3 (1.2)
3	37.9 (5.9)	.210	7.9 (1.5)
4	15.0 (5.3)	.191	3.8 (0.8)
5	17.8 (8.7)	.125	4.9 (1.6)

Significant ANOVA Results

Effect	η^2
Age	.256
Class	.071
Ethnicity	.013
Gender	.006
Class × Gender	.050

Differences in Percentages Between Gender & Differences Within Ethnic Groups

Latent	Gender		Within Ethnicity		
Class	Males	Females	A-A	C	Other
1	57.6	42.4	34.3	26.0	20.0
2	27.9	72.1	20.3	20.4	41.8
3	55.1	44.9	17.5	20.7	10.9
4	46.6	53.4	21.7	18.8	20.0
5	66.4	33.6	6.3	14.1	7.3

Other Conclusions About Latent Classes

- Age effects are clear. Class 1 is oldest, followed by Classes 3 & 2. Class 4 is youngest, preceeded by Class 5.
  - This appears to mirror the score differences as well.
- Class 1 has largest percent of A-A and C members.
- Males tended to predominate in Classes 1, 3, and 5, and higher percentages of females were in Classes 2 and 4.

Some Conclusions for Study 3

- Examinees separated along a secondary nuisance dimension appeared to differ on measures of motor behavior.

- Use of manifest categories is probably misleading and contributes little to understanding what may have caused DIF.

What we can learn

- Mixture Rasch models used to discretize a secondary nuisance dimension to provide a useful tool for better understanding the differences between examinees advantaged and disadvantaged by the DIF items.
- Identifying the differences in propensities of members of each latent class can help to identify the nuisance dimension that might be causing the DIF.

Comparison of Latent Class Propensities

HARD

1. locomotor position
2. object control mechanics
3. positioning off-ground of feet, arm, and body
4. positioning trunk or arms
5. proper position (arm, elbow, foot, leg)

EASY

1. object control, mechanics
2. locomotor, off ground
3. arm movement, weight transfer
4. locomotor, running, galloping, sliding or correct position for jumping, striking, or catching
5. rotation of body, weight transfer