RICHARD SAWATZKY[1,2]
BRUNO ZUMBO[3]
TOLULOPE SAJOBI[4]
LARA RUSSELL[2]
ANNE GADERMANN[3]
JACEK KOPEC[3]
JUXIN LIU[5]
PAMELA RATNER[3]
AMERY WU[3]
LISA M. LIX[6]

Exploring the potential of a mixture-computerized adaptive test for use in heterogeneous populations

PROMIS®: Global Advances in Methodology and Clinical Science

[1]TRINITY WESTERN UNIVERSITY, [2]CENTRE FOR HEALTH EVALUATION AND OUTCOME SCIENCES, [3]UNIVERSITY OF BRITISH COLUMBIA, 4UNIVERSITY OF CALGARY, [5]UNIVERSITY OF SASKATCHEWAN, [6]UNIVERSITY OF MANITOBA

Dublin, Ireland
October, 2018

Background

CATs in heterogeneous population

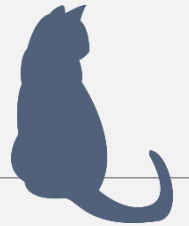
People may interpret and respond to PRO questions in systematically unique ways because of:

- Demographic and cultural differences
- Different health experiences, life circumstances, personality

In this situation, PRO scores could be biased and not directly comparable across different individuals or groups

Background

Computer Adaptive Tests (CATs)



Requirement

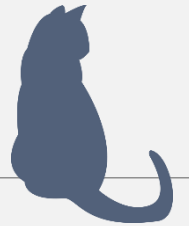
- Invariant IRT-calibrated measurement model parameters that are applicable to all individuals in the target population

Statistical condition: Local independence

- Exchangeable items
- Exchangeable sampling units

Potential solution:

Mixture CAT to accommodate heterogeneity



Goal

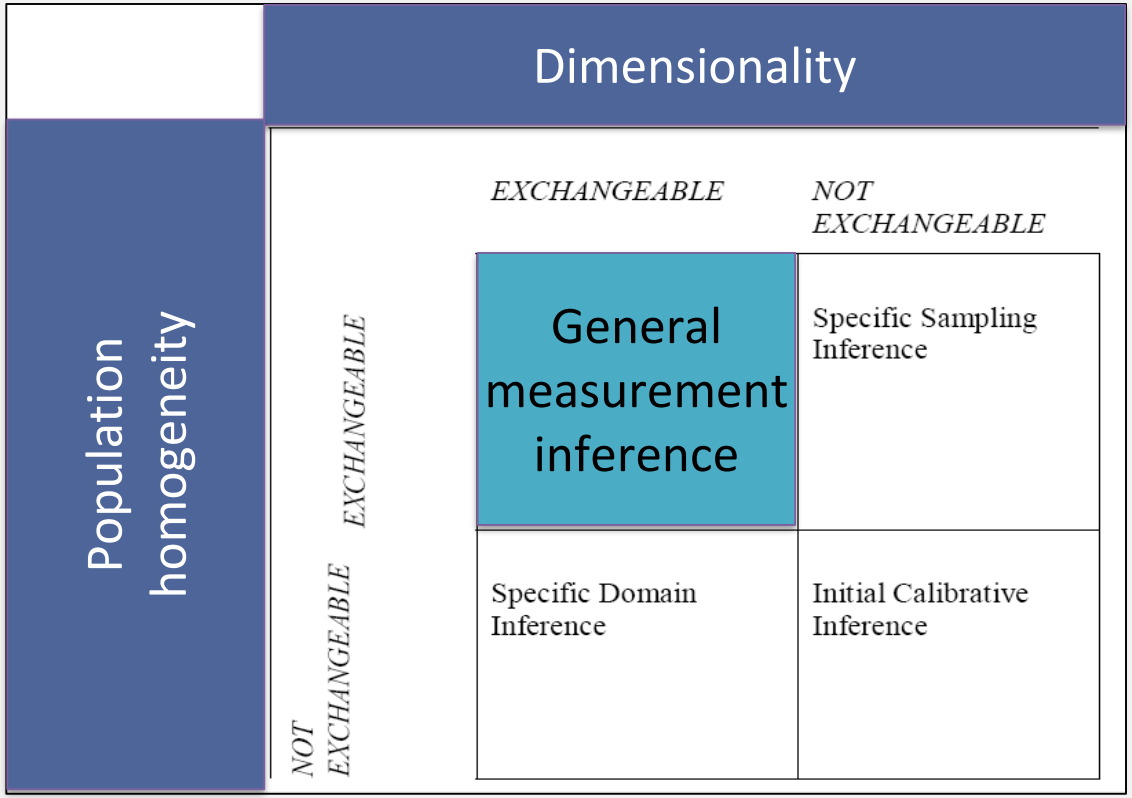
- Develop CAT scoring algorithms that adjust for heterogeneity

Research aims

- Examine implications of population heterogeneity:
  1. Accuracy of CAT scores (extent of bias)
  2. Efficiency and coverage of item selection
  3. Sensitivity in detecting longitudinal change and individual/group differences

Theoretical foundations

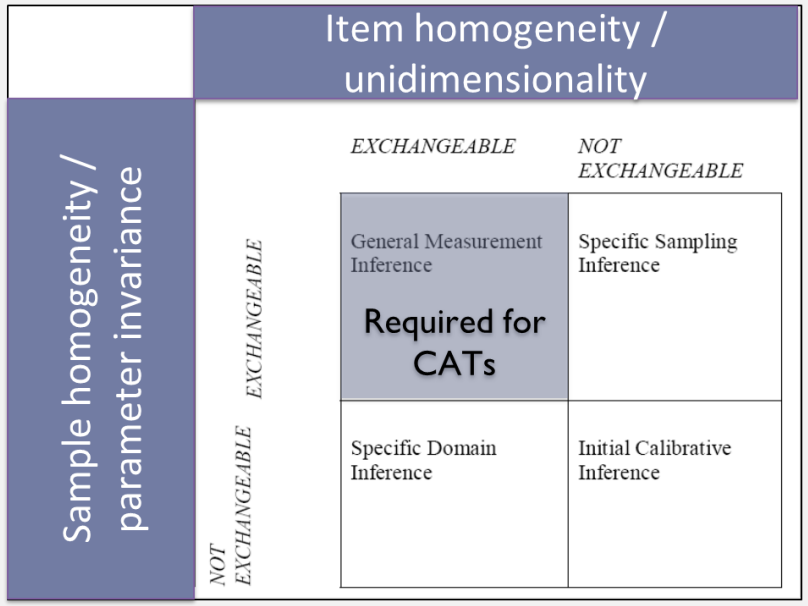
Zumbo's Draper-Lindley-de Finetti (DLD) framework



Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26: Psychometrics, pp. 45-79). Amsterdam: Elsevier Science.

Application to CATs

Computer adaptive testing requires two conditions for “general measurement inference”:



1. Item homogeneity / unidimensionality

- The items must be exchangeable so that the scores of individuals who answered different questions are comparable on the same scale.

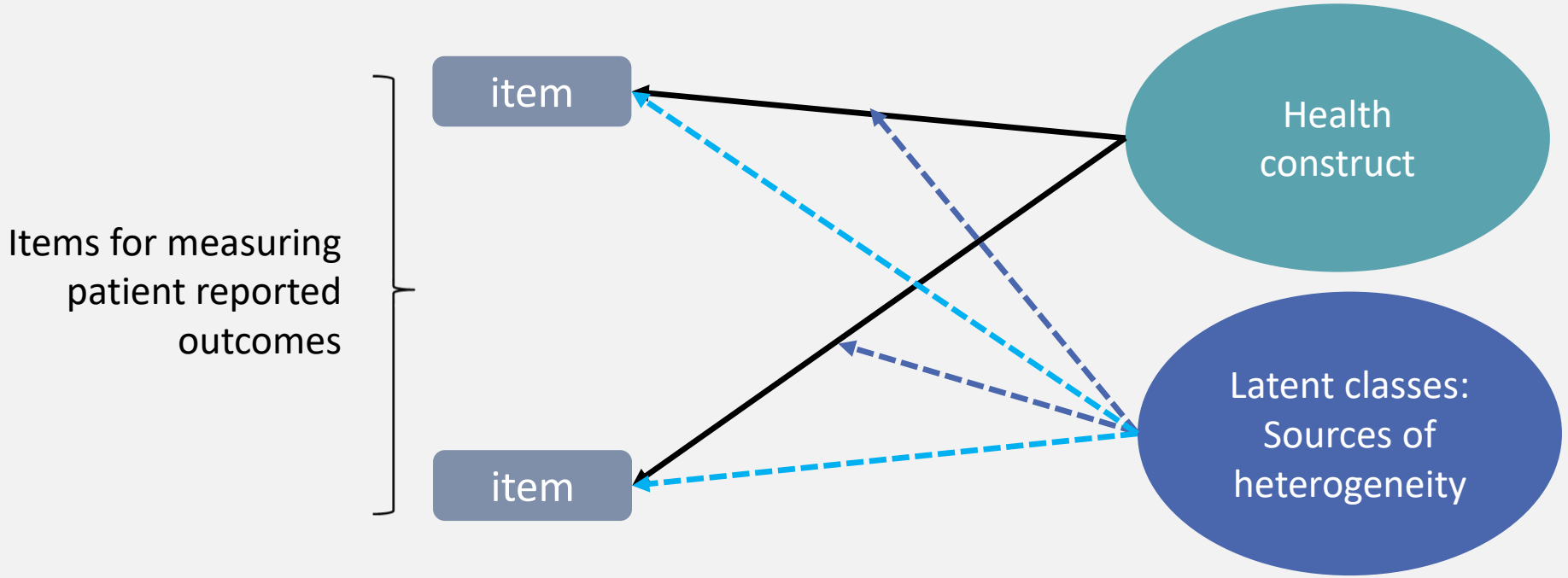
2. Sample homogeneity / parameter invariance

- The sampling units must be exchangeable (the items’ parameters must be invariant) so that the scores are comparable irrespective of any differences among individuals other than the characteristic being measured

Mixture CAT

Based on latent variables mixture models (LVMM)

Item difficulty and discrimination parameters vary across latent classes that represent heterogeneity in the population.



Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research, 21(4), 637-650. doi: 10.1007/s11136-011-9976-6*

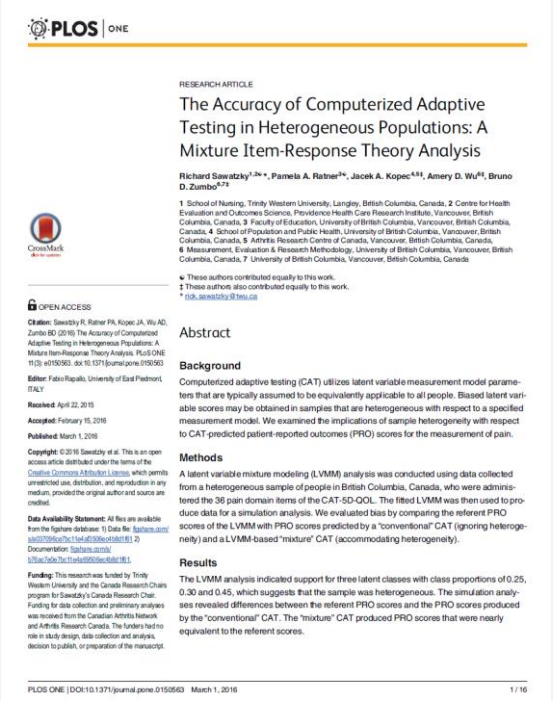
Mixture CAT

Accommodating population heterogeneity

Predict latent class membership

Administer LVMM-calibrated items (weighted by probability of class membership)

Calculate latent-class adjusted theta and information



PLOS ONE

RESEARCH ARTICLE

The Accuracy of Computerized Adaptive Testing in Heterogeneous Populations: A Mixture Item-Response Theory Analysis

Richard Sawatzky[1,2☯*], Pamela A. Ratner[3☯], Jacek A. Kopec[4,5‡], Amery D. Wu[6‡], Bruno D. Zumbo[6,7‡]

1 School of Nursing, Trinity Western University, Langley, British Columbia, Canada, 2 Centre for Health Evaluation and Outcomes Science, Providence Health Care Research Institute, Vancouver, British Columbia, Canada, 3 Faculty of Education, University of British Columbia, Vancouver, British Columbia, Canada, 4 School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada, 5 Arthritis Research Centre of Canada, Vancouver, British Columbia, Canada, 6 Measurement, Evaluation & Research Methodology, University of British Columbia, Vancouver, British Columbia, Canada, 7 University of British Columbia, Vancouver, British Columbia, Canada

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* rick.sawatzky@twu.ca

OPEN ACCESS

Abstract

Background

Computerized adaptive testing (CAT) utilizes latent variable measurement model parameters that are typically assumed to be equivalently applicable to all people. Biased latent variable scores may be obtained in samples that are heterogeneous with respect to a specified measurement model. We examined the implications of sample heterogeneity with respect to CAT-predicted patient-reported outcomes (PRO) scores for the measurement of pain.

Methods

A latent variable mixture modeling (LVMM) analysis was conducted using data collected from a heterogeneous sample of people in British Columbia, Canada, who were administered the 36 pain domain items of the CAT-5D-QOL. The fitted LVMM was then used to produce data for a simulation analysis. We evaluated bias by comparing the referent PRO scores of the LVMM with PRO scores predicted by a "conventional" CAT (ignoring heterogeneity) and a LVMM-based "mixture" CAT (accommodating heterogeneity).

Results

The LVMM analysis indicated support for three latent classes with class proportions of 0.25, 0.30 and 0.45, which suggests that the sample was heterogeneous. The simulation analyses revealed differences between the referent PRO scores and the PRO scores produced by the "conventional" CAT. The "mixture" CAT produced PRO scores that were nearly equivalent to the referent scores.

PLOS ONE | DOI:10.1371/journal.pone.0150563 March 1, 2016 1 / 16

Sawatzky, R., Ratner, P. A., Kopec, J. A., Wu, A. D., & Zumbo, B. D. (2016). The Accuracy of Computerized Adaptive Testing in Heterogeneous Populations: A Mixture Item-Response Theory Analysis. *PLoS One, 11(3), e0150563.*

Objectives

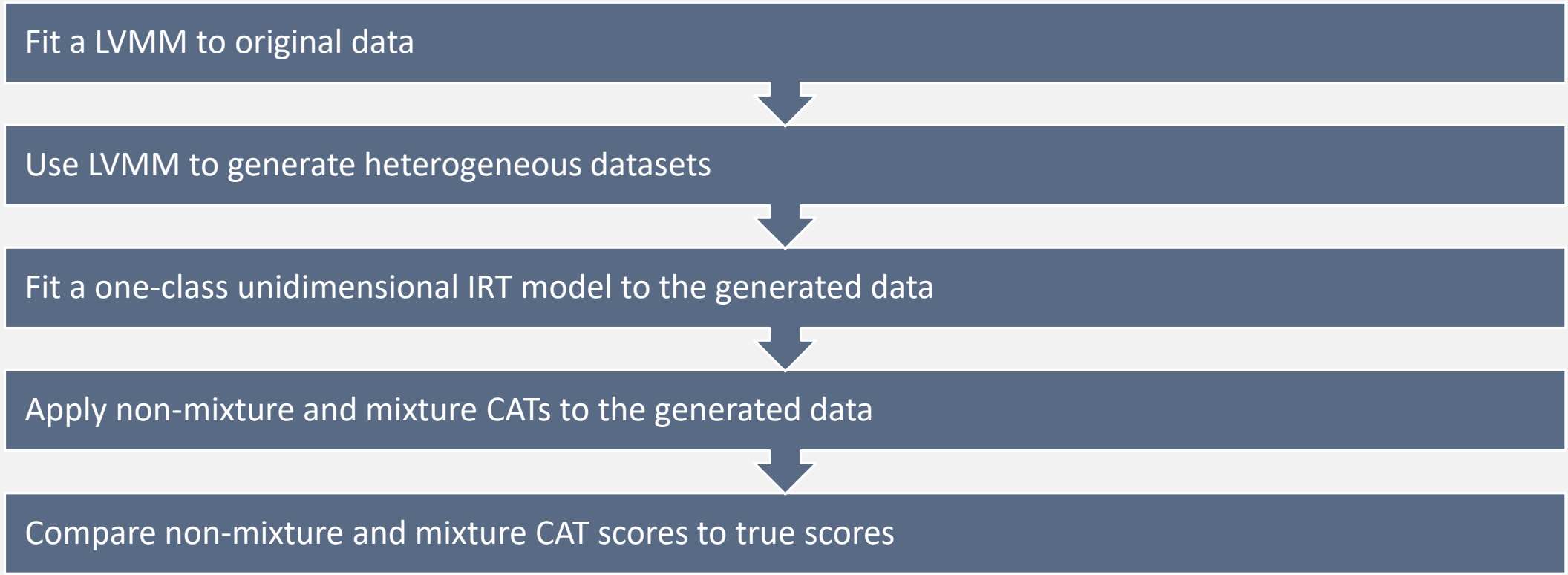
To examine the potential of using latent variable mixture models (LVMMs) to estimate heterogeneity-adjusted “mixture CAT” scores

To compare mixture CAT and non-mixture CAT scores to true scores.

Methods

Methods

IRT mixture model simulation study based on item parameters obtained from real data.



Fit a LVMM to original data

Item bank measuring **daily activities**

- 39 items measuring the ability to perform common daily activities
- One of the item banks of the CAT-5D-QOL (Kopec et al., 2006)

Sample

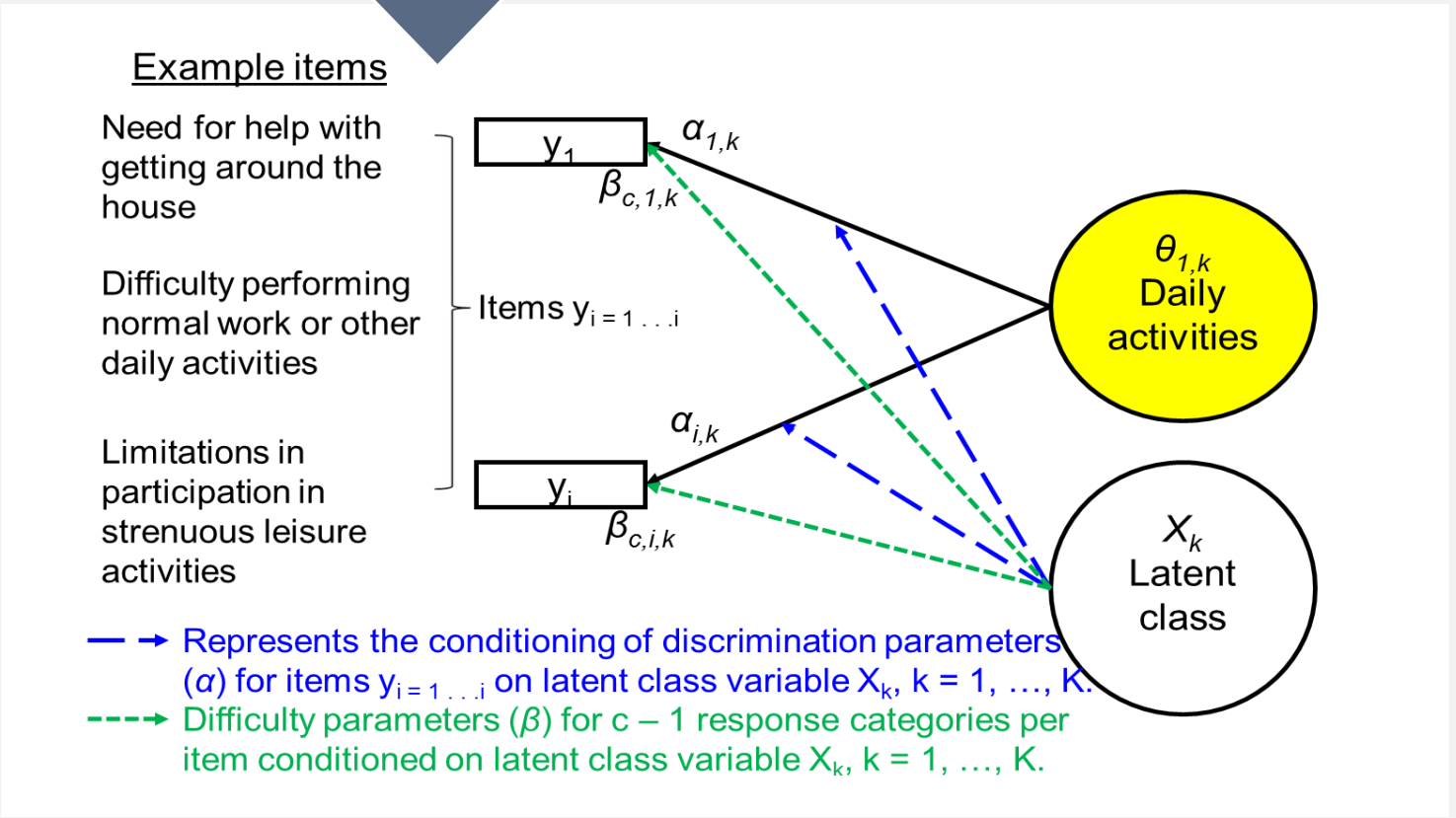
- Adults from two rheumatology clinics (N = 340)
- Adults on a joint replacement surgery waiting list (N = 331)
- Stratified random community sample (N = 995)

Statistical model

- A 2-class mixture of Samejima's 2-parameter Graded Response Model (GRM)

Fit a LVMM to original data

MIXTURE OF THE GRADED RESPONSE MODEL



Generate heterogeneous data

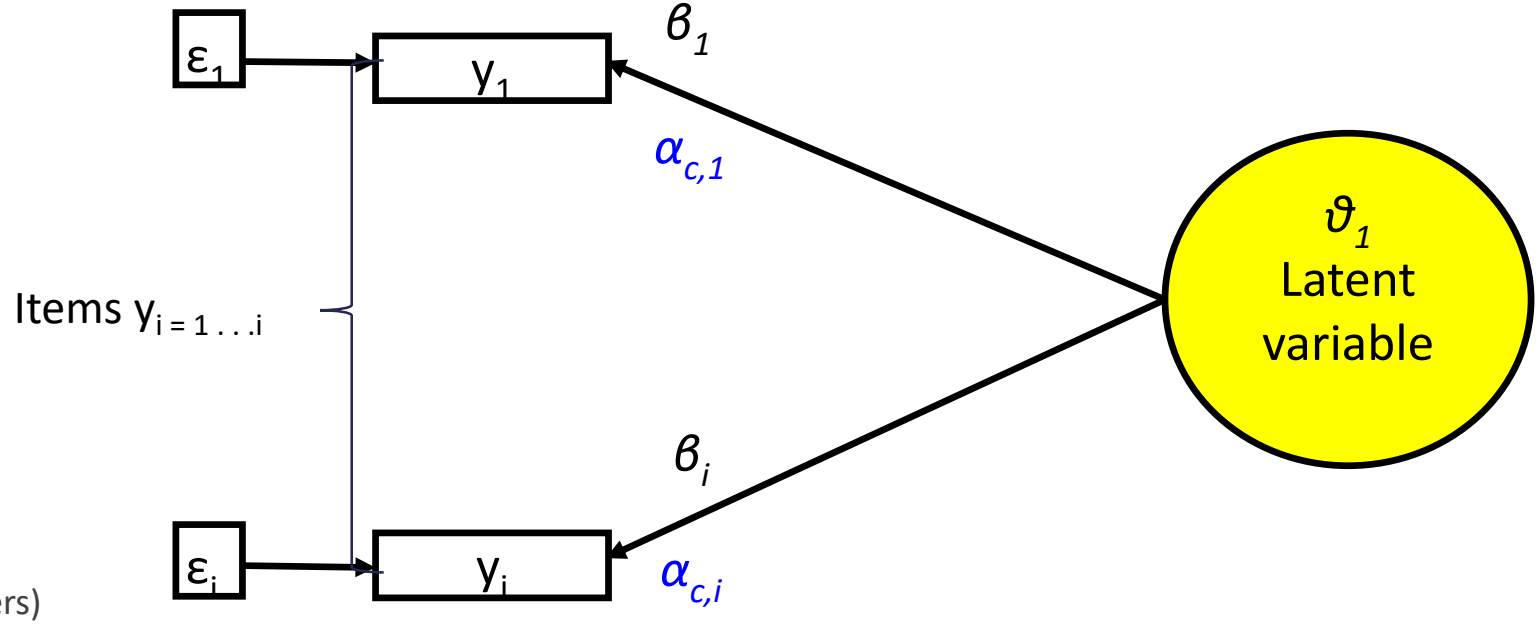
100 datasets (of N=1,000 each) were generated using

- Parameters based on the IRT mixture model
- Randomly-generated normally-distributed “true-theta scores”

Fit a one-class IRT model to the generated data

- 2-parameter GRM (ignores sample heterogeneity)
- Parameter estimates and predicted scores were saved.

Fit a one-class IRT model to the generated data



β = slopes
(a.k.a. discrimination parameters)

α = thresholds
(a.k.a. difficulty parameters)

Apply and compare non-mixture and mixture CAT

Apply CAT to the generated data	• Mixture CAT (based on LVMM parameters) • Conventional CAT (based on 1-class GRM parameters)
CAT stopping rules	• Standard error <= 0.20 • Maximum number of items: 10
Saved data	• The items that were applied • The CAT-predicted theta scores for each individual • The CAT-predicted information for each individual
Results	• Comparison of mixture and non-mixture CAT scores to true scores

Results

Global fit of the LVMM model

Model	*P*	BIC	LL ratio[1]	Entropy	Class proportions[2]	
					Class 1	Class 2
1 class	192	76390.31				
2 classes	385	74064.00	3143.76	0.84	0.65	0.35

N = 1,662. *P* = number of model parameters. BIC = Bayesian Information Criterion. LL = log likelihood
[1] Likelihood ratio of 1 and 2 class models. Statistical significance was confirmed using a bootstrapped likelihood ratio test with simulated data. [2] Based on posterior probabilities.

- A relative improvement in model fit was obtained when 2 classes were specified.
- The sample is not homogeneous with respect to a unidimensional structure for the daily activities items.

Description of latent classes

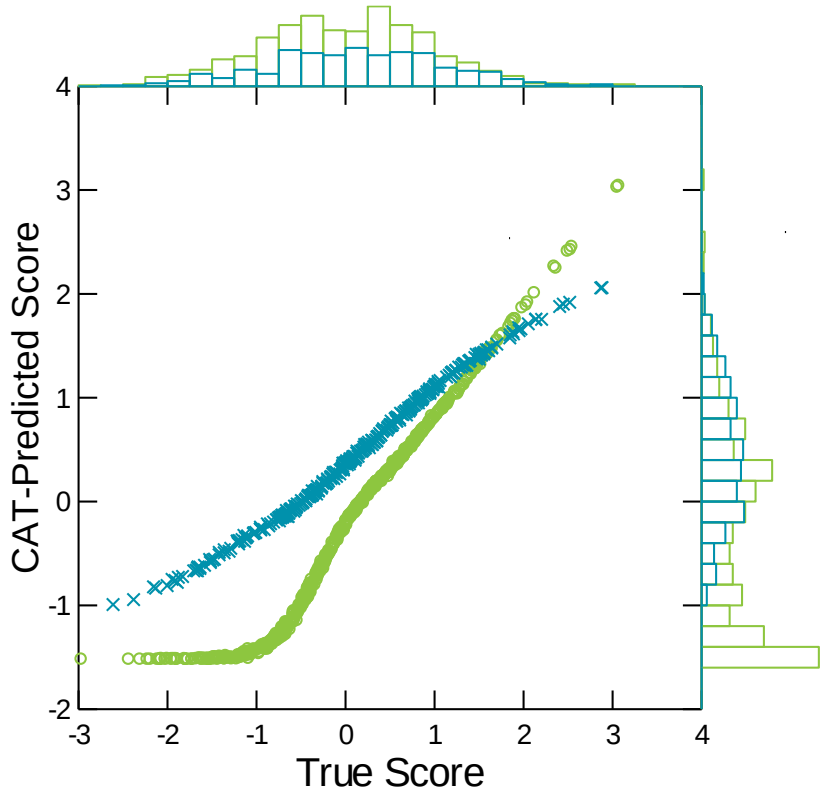
Variables	Class 1	Class 2	OR (95% CI)
Gender = female	59.7%	62.3%	1.0 (0.8-1.3)
Age (10 yr increments) (means(sd))	5.3(1.6)	**6.3(1.4)**	0.8 (0.7-0.8)*
Has a medical problem	78.9%	**96.0%**	2.6 (1.5-4.5)*
Has osteo-arthritis	26.0%	**56.3%**	2.2 (1.7-2.7)*
Has rheumatoid arthritis	24.2%	65.1%	1.1 (0.8-1.5)
Uses one medication	24.8%	34.9%	1.3 (0.9-2.0)
Uses two or more medications	45.8%	**69.7%**	1.6 (1.1-2.4)*
Has been hospitalized during the past year	16.7%	27.6%	1.2 (0.9-1.6)
Self-reported health status (1 = excellent; 5 = very poor) (mean (sd))	2.5(1.1)	**2.9(1.0)**	1.1 (1.0-1.3)*

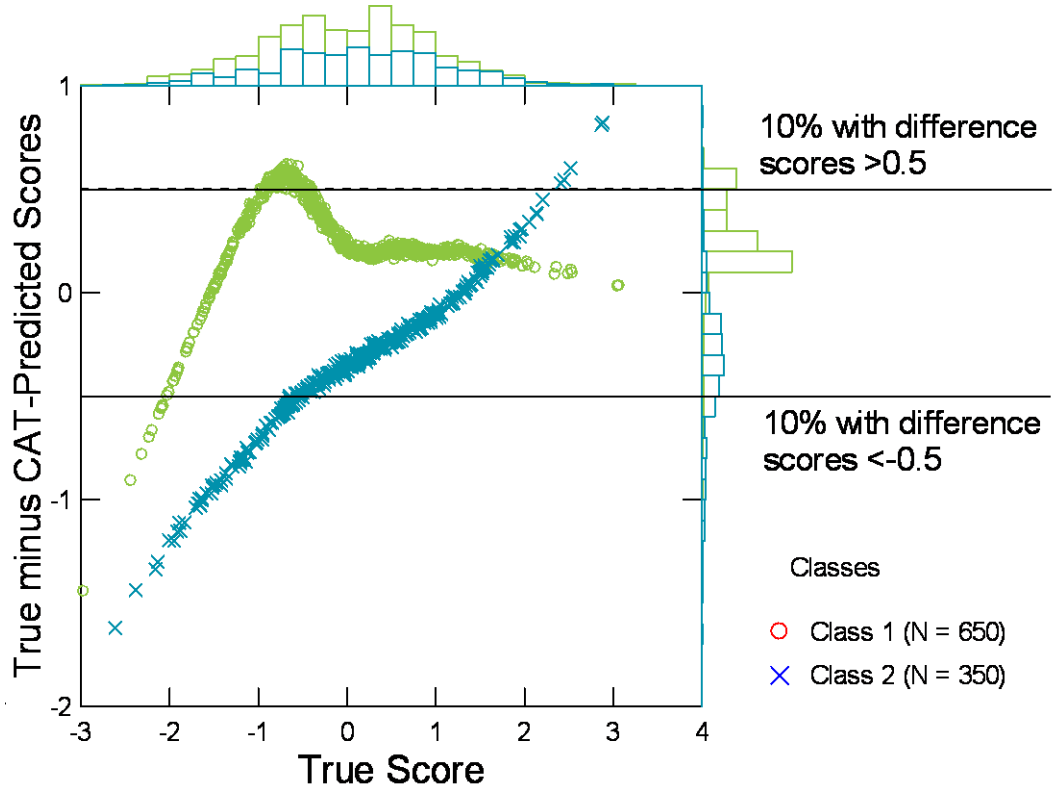
Nagelkerke's R^2 = 21.0%. OR = adjusted odds ratio comparing class 2 to class 1. *p>0.05.

People in class 2 are relatively older, more likely to have a chronic condition, and more likely to use two or more medications.

CAT-predicted scores

Ignoring heterogeneity (based on 1-class GRM)

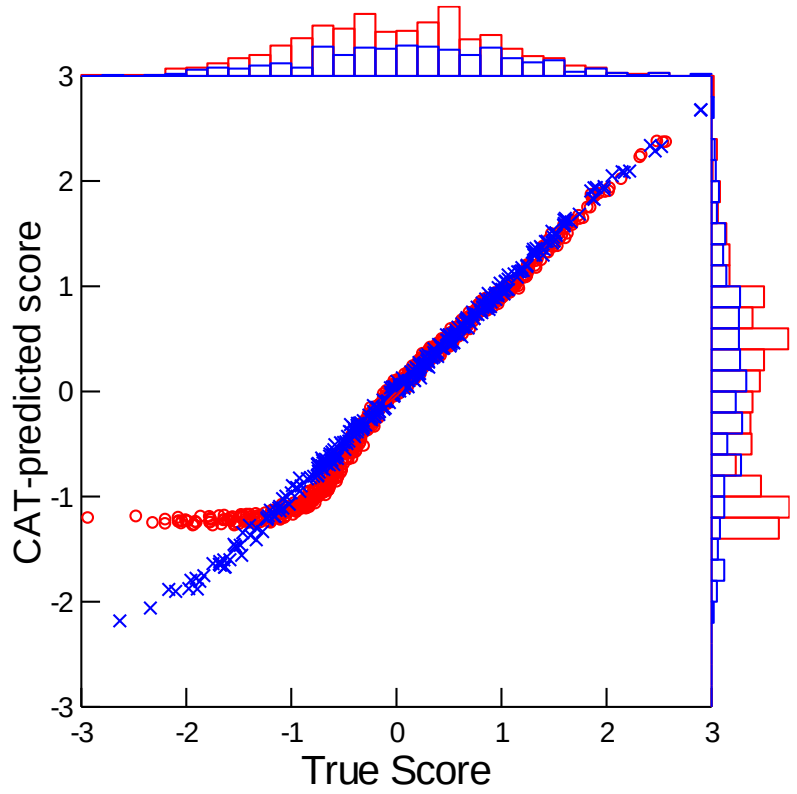


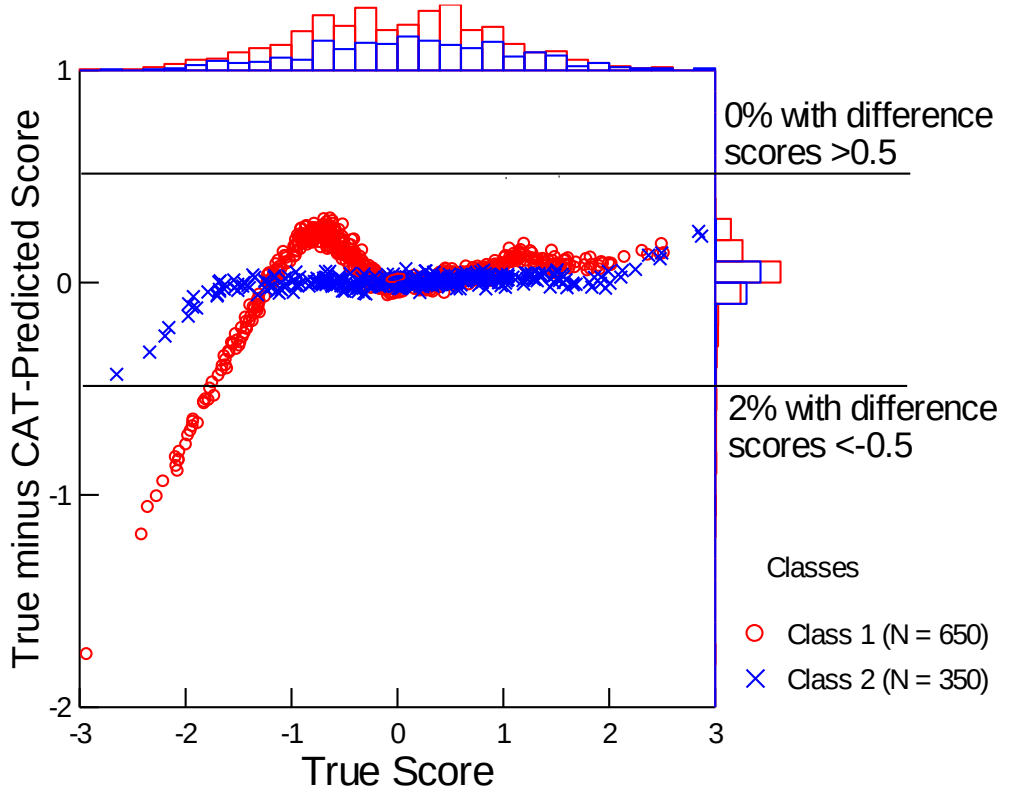


- 20% of the predicted scores were off by $\pm$ 0.5 standard deviations from the true scores

Mixture CAT-predicted scores

Adjusting for heterogeneity (based on 2-class LVMM)





- Only 2% of the predicted scores were off by ± 0.5 standard deviations from the true scores

Discussion

The challenge of population heterogeneity

People may not interpret and respond to questions about their health and quality of life in the same way.

Unaccounted for heterogeneity could be a source of measurement error.

Potential solution

Mixture CATs based on LVMMs could be used to adjust PRO scores in heterogeneous populations, leading to improved:

- accuracy of CAT-predicted PRO scores
- efficiency and coverage of item selection

Discussion

Limitations

Mixture CATs require reliable prediction and replication of latent classes

Current and ongoing research

1. Real data and simulation studies on latent class prediction
2. LVMM calibration of existing item banks
3. Comparative evaluation of mixture versus non-mixture CATs:
   - Efficiency and coverage of item selection
   - Sensitivity in detecting longitudinal change and individual/group differences

Thank you!

Richard Sawatzky, RN, PhD

PROFESSOR & CANADA RESEARCH CHAIR IN PERSON-CENTRED OUTCOMES
SCHOOL OF NURSING, TRINITY WESTERN UNIVERSITY;
CENTRE FOR HEALTH EVALUATION AND OUTCOMES SCIENCES, PROVIDENCE HEALTH CARE

Rick.Sawatzky@twu.ca

TRINITY WESTERN UNIVERSITY

CHÉOS Centre for Health Evaluation & Outcome Sciences

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada