MEASURING PATIENT-REPORTED OUTCOMES IN DIVERSE POPULATIONS:

Novel Methods for Equitable People-Centred Health Measurement

Richard Sawatzky, RN, PhD

Professor & Canada Research Chair in Person-Centred Outcomes
School of Nursing, Trinity Western University, Langley BC

Centre for Health Evaluation and Outcomes Sciences,
St. Paul's Hospital, Vancouver, BC

Salgrenska Academy, Centre for Person-Centred Care,
University of Gothenburg, Sweden

CHÉOS
Centre for Health Evaluation & Outcome Sciences

TRINITY WESTERN UNIVERSITY

Measuring PROs in diverse populations

- Introduction: Is it defensible to believe that all people interpret and respond to questions for measuring PROs in the same way? Why is this of concern?
- Possible solution: How can latent variable mixture models (LVMMs) be used to measure PROs in diverse populations?
- Example applications of LVMMs:
  1. Explain heterogeneous responses to PRO measurement items
  2. Identify invariant PRO measurement items
  3. Adjust scoring and selection of items in computerized adaptive tests

Acknowledgements

Statistical Methods for Patient-Reported Outcome Measures

- Principal investigators: R. Sawatzky, T.T. Sajobi.
- Researchers: E.R. Bohm, S. Bryan, A.M. Gadermann, W. Hopman, J. Liu, L. Lix, L. Russell, N.E. Mayo, K.M. Sibley, X. Wu., B.D. Zumbo. Knowledge users: L.N. Cuthbertson, G.K. Webster.

No one size fits all for measuring what matters: Supporting the use of PRO data in healthcare

- Principal Investigators: R. Sawatzky & J.-Y. Kwon
- Researchers: L.M. Lix, T.T. Sajobi, K. Sibley, L. Russell. Knowledge user team members: R. Brahmbhatt, L. Cuthbertson, S. Lauck, A. Tecson

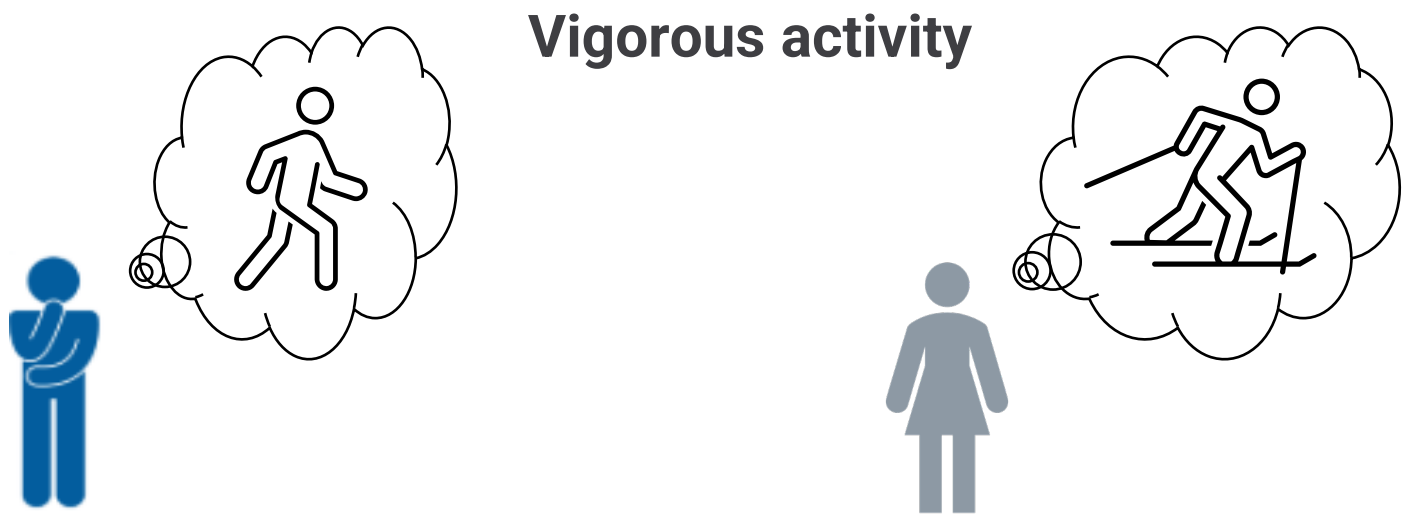
This work was undertaken, in part, thanks to funding from the Canadian Institutes of Health Research, Michael Smith Foundation for Health Research, Canada Research Chairs Program

Introduction

- Is it defensible to believe that all people interpret and respond to questions for measuring PROs in the same way?
- Why is this of concern?

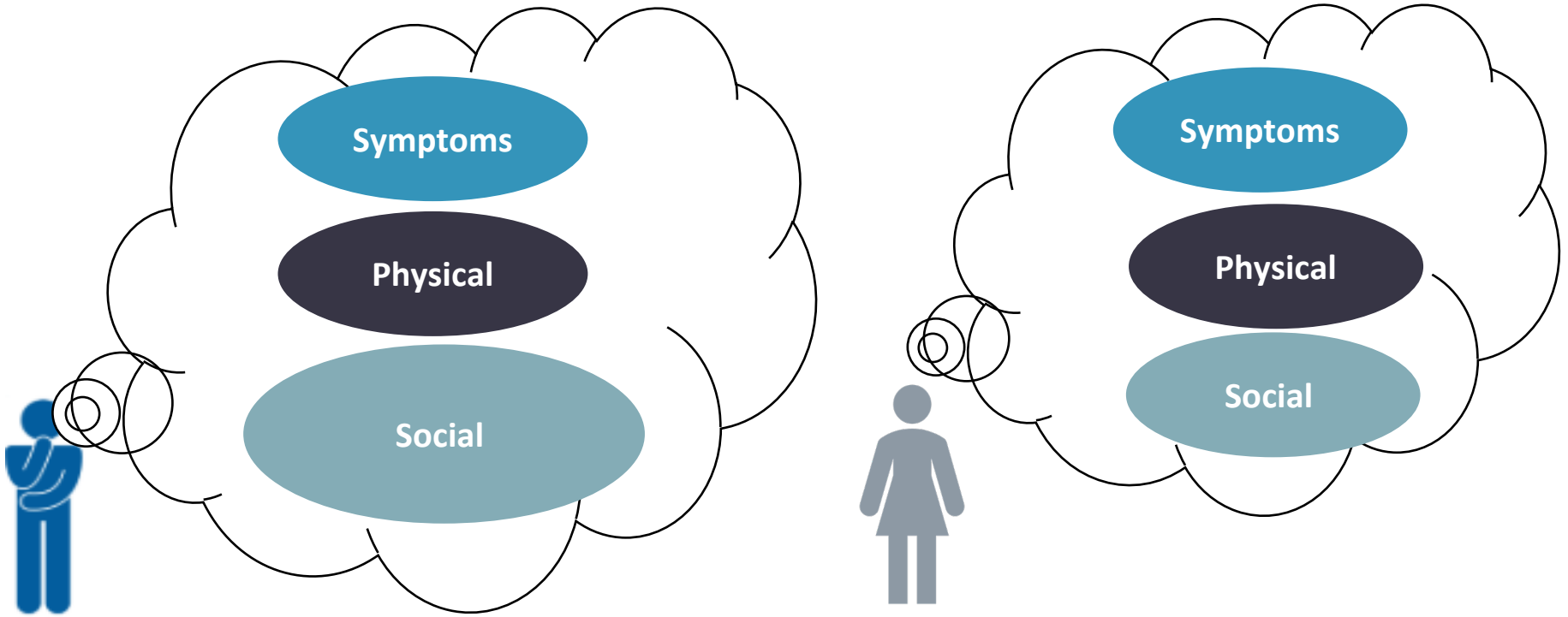
Measuring PROs in heterogeneous populations

- Do questions for measuring physical activity have the same meaning for different people?



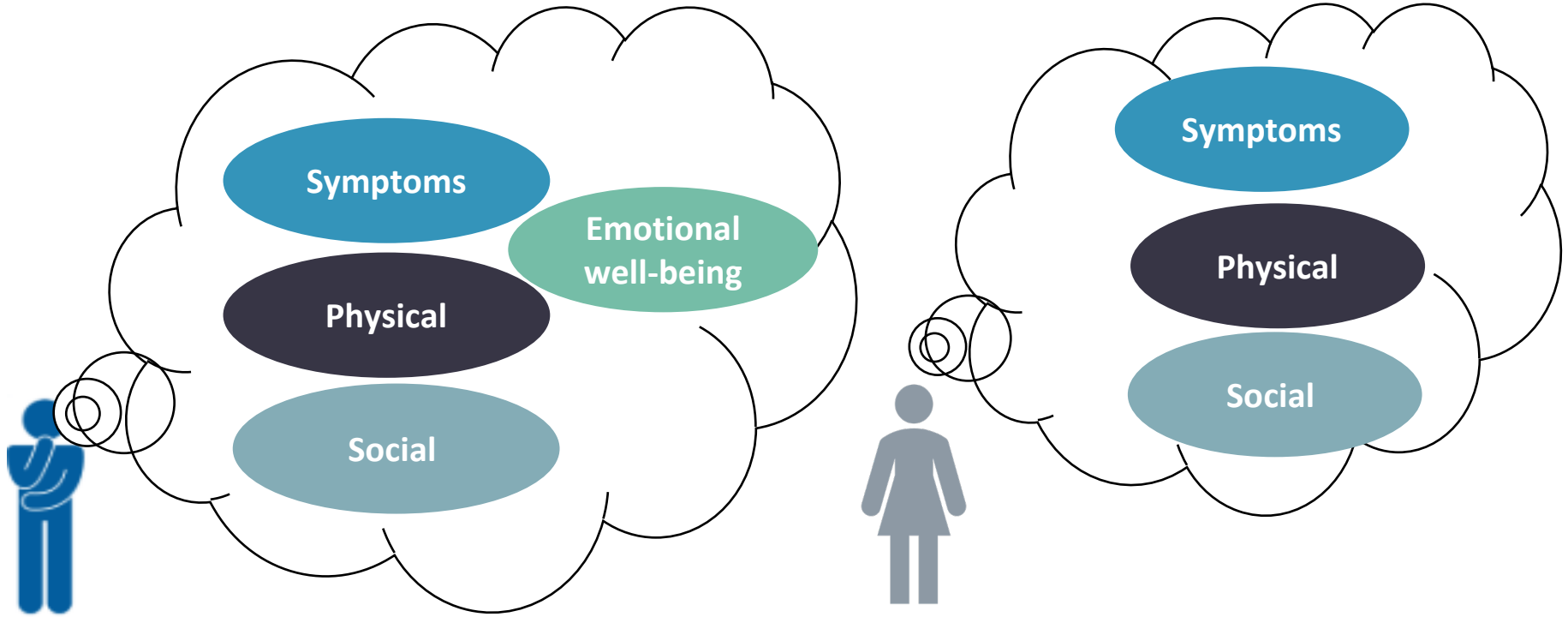
Measuring PROs in heterogeneous populations

- Are the same health domains equally important for different people?



Measuring PROs in heterogeneous populations

- Are different dimensions of health relevant for some people and not for others?



Measuring PROs in heterogeneous populations

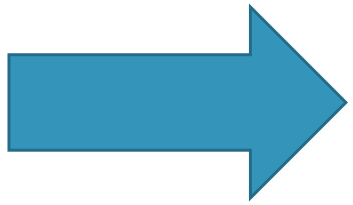
- Does a change in PRO scores have the same meaning for different people?

Measuring PROs in heterogeneous populations

There are many factors that may influence how people interpret and respond to questions for measuring PROs

- Social determinants of health
- Health history
- Healthcare experiences
- Personality

Measuring PROs in diverse populations



Ignoring individual differences in how people interpret and respond to questions for measuring their health could lead to measurement biases that reinforce inequities by hiding or misrepresenting health outcomes of some people.

What are QOL Assessments? | Applications in Healthcare | Measurement Validity | About/Contact

Measurement Validity

Learn about measurement validity, approaches and methods for using patient-reported outcomes (PROs) to assess change and diverse people.

Introductory Video and Webinar

Whiteboard Video introducing:

1. Use of PRO measures
2. Response shift
3. Differential item functioning

ISOQOL Webinar on interpretation of PROs in clinical practice: Solutions for assessing change and diverse people

Recorded webinar and slides are freely to everyone

This research was undertaken, in part, thanks to funding from the Canadian Institutes of Health Research, the Michael Smith Foundation for Health Research, and the Canada Research Chairs Program in support of Dr. Richard Sawatzky's Research Chair in Person-Centred Outcomes at Trinity Western University.


Current Oncology

Article

Patient-Reported Outcomes Measurement in Radiation Oncology: Interpretation of Individual Scores and Change over Time in Clinical Practice

Jae-Yung Kwon [1,2,*], Lara Russell [3], Theresa Coles [4], Robert J. Klaassen [5], Kara Schick-Makaroff [6], Kathryn M. Sibley [7,8], Sandra A. Mitchell [9] and Richard Sawatzky [3,10,11]

Abstract: Tools for measuring patients' perceived health and quality of life, such as patient-reported outcome measures (PROMs), inform clinical decisions for patients requiring radiation therapy. However, there may be inconsistencies in how patients interpret and respond to PROMs due to cultural, environmental, personal, or experiential factors. Differential item functioning (DIF) and response shift (RS) refer to differences in the meaning of PROMs between patients or over time (respectively). DIF and RS can threaten the accurate interpretation and use of PROMs, potentially resulting in erroneous conclusions about effectiveness, and flawed individual-level clinical decision-making. Given the empirical evidence of DIF and RS, we aim to review clinical implications and solutions for addressing DIF and RS by providing vignettes from collaborative examinations with workshop participants, as well as the literature. By making these methodological concepts accessible and relevant, for practice, clinicians may feel more confident to ask clarifying questions of patients when PROM scores and the contextual patient information do not align. PROM scores need to be interpreted via dialogue with the patient to avoid misinterpretation due to DIF and RS, which could diminish patient–clinician communication and impede shared decision-making. This work is part of an interdisciplinary knowledge translation initiative focused on the interpretation of PROM scores by clinically-oriented audiences.

Keywords: patient-reported outcome measures; differential item functioning; response shift

Kwon, J.-Y., Russell, L., Coles, T., Klaassen, R. J., Schick-Makaroff, K., Sibley, K. M., Mitchell, S. A., & Sawatzky, R. (2022). Patient-Reported Outcomes Measurement in Radiation Oncology: Interpretation of Individual Scores and Change over Time in Clinical Practice. Current Oncology, 29(5), 3093-3103. 10.3390/curroncol29050251

Learn About Analytical Methods

Independent online learning module

Includes a basic introduction with expository analyses and syntax for:

- **Differential item functioning methods** to compare different people
- **Response shift methods** to assess change over time

Research & Publications by Our Team

- Practical implications
- Theoretical foundations
- Analytical methods

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

MICHAEL SMITH FOUNDATION FOR HEALTH RESEARCH
BC's health research funding agency

TRINITY WESTERN UNIVERSITY

CONTACT info@healthyqol.com

Possible solutions

- How can latent variable mixture models (LVMMs) be used to investigate population heterogeneity with respect to the measurement of PROs?

Investigating measurement bias

Measurement biases within PROMs have been extensively investigated under the headings of ‘Differential Item Functioning’ and ‘Measurement Invariance’

MEASUREMENT INVARIANCE (MI)

measurement properties are equivalent for all people

DIFFERENTIAL ITEM FUNCTIONING (DIF)

the same item may not equivalently reflect the outcome being measured when it is applied to different people

DIFFERENTIAL TEST FUNCTIONING

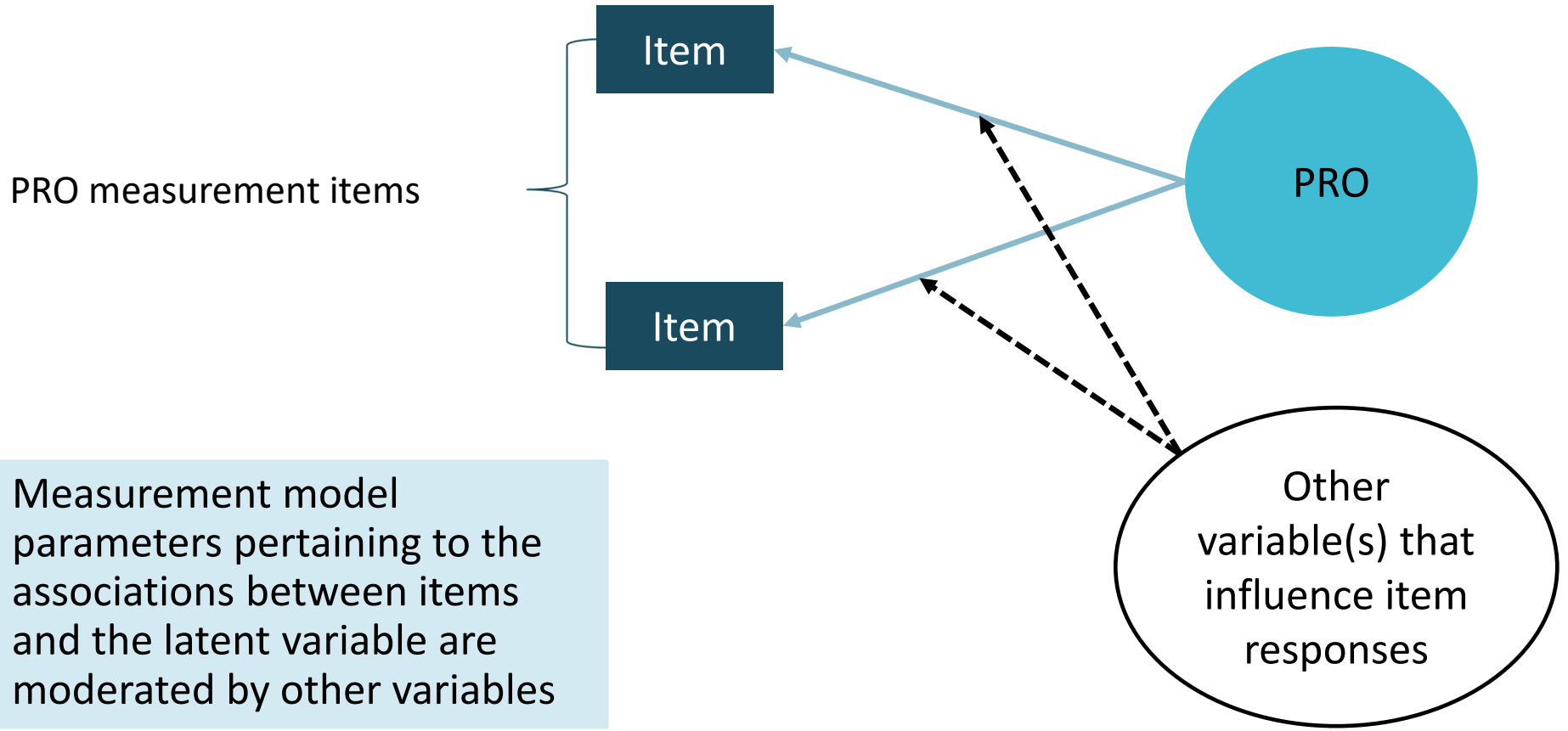
the same PROM summary scores are not equivalently representative of the measured health outcome for different people

PSYCHOMETRIC TOOLBOX

- **Multi-group confirmatory factor analysis**
- **Multi-group item response theory / RASCH**
- **Logistic regression differential item functioning analysis**
- **Multiple Indicators Multiple Causes (MIMIC) analysis**

Investigating measurement invariance

Examine heterogeneity using latent variable mixture models (LVMMs)



Measurement model parameters pertaining to the associations between items and the latent variable are moderated by other variables

Measurement invariance research

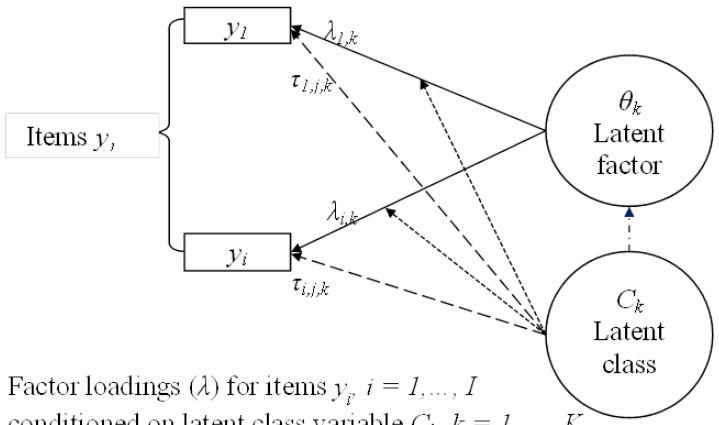
Previous research

Focus on a prior identified sources of measurement biases

- **Socio-demographic differences:** age, sex, gender, education, language, culture, ethnicity, employment status, socio-economic status
- **Health status indicators:** multimorbidity, social deprivation, body weight, and clinical vs. non-clinical status

Moving forward

- Emphasis on intersectionality of individual differences within heterogeneous populations
- Identify individual differences, and interactions among them, that are not identified a priori



Factor loadings (λ) for items y_i, $i = 1, \ldots, I$ conditioned on latent class variable C_k, $k = 1, \ldots, K$.

Thresholds (τ) for $j - 1$ response categories per item conditioned on latent class variable C_k, $k = 1, \ldots, K$.

Variance of the latent factor (θ) conditioned on latent class variable C_k, $k = 1, \ldots, K$.

We will not discuss the mathematic details today!

Latent variable mixture models: a promising approach for the validation of patient reported outcomes

Richard Sawatzky · Pamela A. Ratner · Jacek A. Kopec · Bruno D. Zumbo

Accepted: 7 July 2011 / Published online: 5 August 2011
© Springer Science+Business Media B.V. 2011

Abstract

Purpose A fundamental assumption of patient-reported outcomes (PRO) measurement is that all individuals interpret questions about their health status in a consistent manner, such that a measurement model can be constructed that is equivalently applicable to all people in the target population. The related assumption of sample homogeneity has been assessed in various ways, including the many approaches to differential item functioning analysis.

Methods This expository paper describes the use of latent variable mixture modeling (LVMM), in conjunction with item response theory (IRT), to examine: (a) whether a sample is homogeneous with respect to a unidimensional measurement model, (b) implications of sample heterogeneity with respect to model-predicted scores (theta), and (c) sources of sample heterogeneity. An example is provided using the 10 items of the Short-Form Health Status (SF-36®) physical functioning subscale with data from the Canadian Community Health Survey (2003) ($N = 7{,}030$ adults in Manitoba).

Results The sample was not homogeneous with respect to a unidimensional measurement structure. Specification of three latent classes, to account for sample heterogeneity, resulted in significantly improved model fit. The latent classes were partially explained by demographic and health-related variables.

Conclusion The illustrative analyses demonstrate the value of LVMM in revealing the potential implications of sample heterogeneity in the measurement of PROs.

Keywords Self-report measurement · Psychometrics · Measurement validity · Physical function

Abbreviations

AIC	Akaike information criterion
BLRT	Bootstrapped likelihood ratio test
BIC	Bayesian information criterion
BIC*	Sample-adjusted Bayesian information criterion
CCHS	Canadian Community Health survey
DIF	Differential item functioning
IRT	Item response theory
GRM	Graded response model
LVMM	Latent variable mixture model
OR	Odds ratio
PRO	Patient reported outcomes
VLMR LRT	Vuong-Lo-Mendell-Rubin likelihood ratio test

Electronic supplementary material The online version of this article (doi:10.1007/s11136-011-9976-6) contains supplementary material, which is available to authorized users.

R. Sawatzky (✉)
School of Nursing, Trinity Western University, 7600 Glover Rd, Langley, BC, Canada
e-mail: rick.sawatzky@twu.ca

P. A. Ratner
School of Nursing, University of British Columbia, 302-6190 Agronomy Road, Vancouver, BC V6T 1Z3, Canada

J. A. Kopec
School of Population and Public Health, University of British Columbia; Arthritis Research Centre of Canada, 895 West 10th Avenue, Vancouver, BC V5Z 1L7, Canada

B. D. Zumbo
ECPS, Measurement, Evaluation and Research Methodology, University of British Columbia, Scarfe Building, 2125 Main Mall, Vancouver, BC, Canada

► The cumulative probability of an item response at or above category *j* within a latent class can be computed as follows:

$$P_{ijk}\left(Y \geq j \mid \theta, C = k\right) = \frac{\exp(-\tau_{ijk} + \lambda_{ik}\theta)}{1 + \exp(-\tau_{ijk} + \lambda_{ik}\theta)}.$$

► Each class has a unique set of parameters that are estimated simultaneously in the latent variable mixture model:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

, where *f* is the mixture of the class-specific distributions, and π_k is the mixing proportion.

► The cumulative probability of an item response at or above category *j* within a latent class can be computed as follows:

$$P_{ijk}\left(Y \geq j \mid \theta, C = k\right) = \frac{\exp(-\tau_{ijk} + \lambda_{ik}\theta)}{1 + \exp(-\tau_{ijk} + \lambda_{ik}\theta)}.$$

► The cumulative probability of an item response at or above category *j* within a heterogeneous population is obtained by:

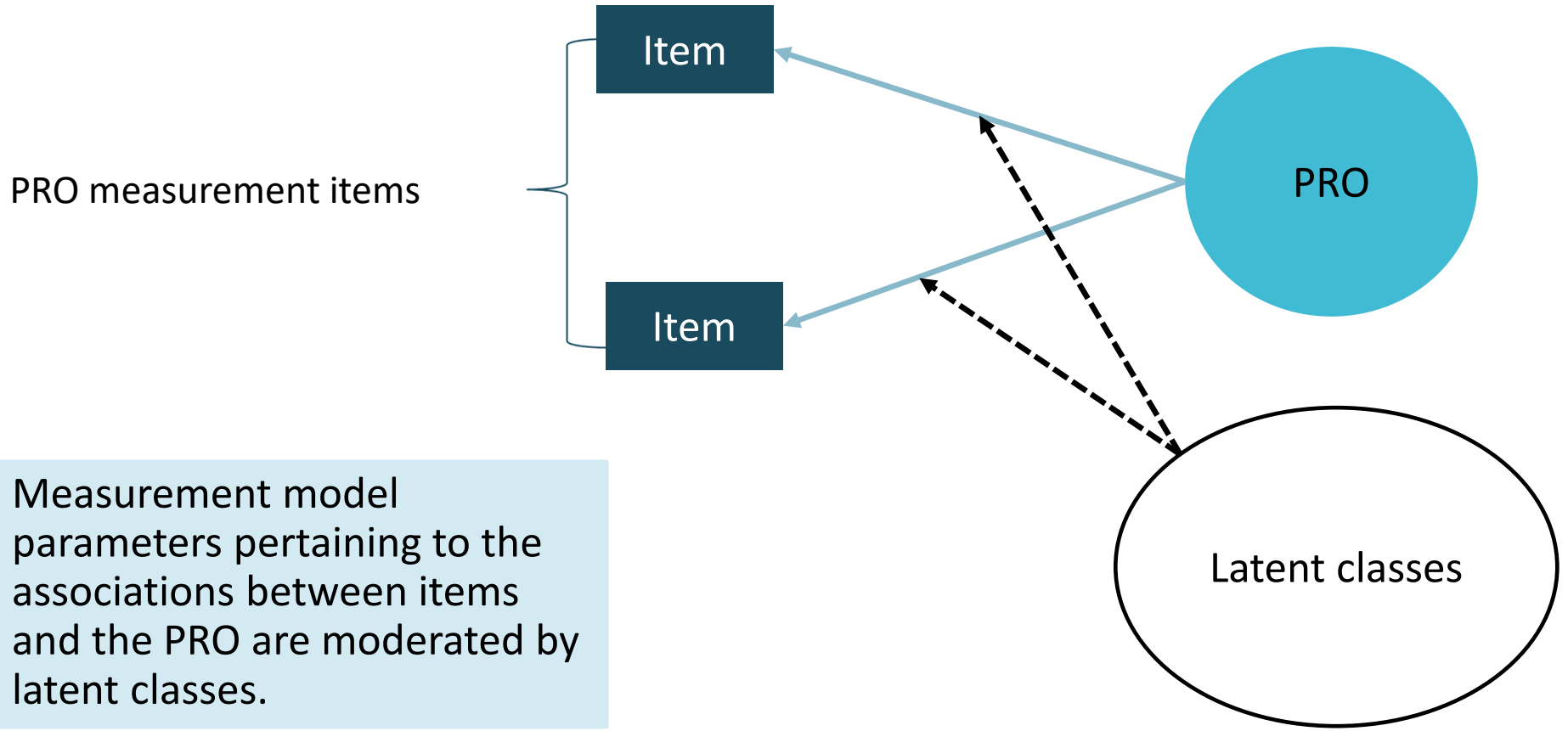
$$P_{ij}(Y \geq j \mid \theta) = \sum_{k=1}^{K} (X_k * P_{ijk}(Y \geq j \mid \theta))$$

, where X_k is the posterior probability of an individual being in class *k*.

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.

Investigating measurement invariance

Examine heterogeneity using latent variable mixture models (LVMMs)



Measurement model parameters pertaining to the associations between items and the PRO are moderated by latent classes.

Analytical approach

Using LVMMs to investigate measurement biases in heterogeneous populations

Establish measurement model	• Use multidimensional item response theory (IRT) to specify the baseline measurement model
Examine heterogeneity	• Specify a latent variable mixture model (LVMM) by adding latent classes to the measurement model.
Assess measurement bias	• Compare scores of a 1-class model (assuming no heterogeneity) and a multi-class model (accommodating heterogeneity)
Explain sources of heterogeneity	• Identify individual differences that predict latent class membership

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.

Three example applications of LVMMs

1. Describe individual differences associated with differential interpretations and responses to PRO measurement items
2. Identify PRO measurement items that are most likely to be invariant within heterogeneous populations
3. Adjust PRO item selection and scoring in computerized adaptive tests (CATs): introducing “mixture-CAT”

Explain heterogeneous responses to PRO measurement items

Richard Sawatzky[1], Lara Russell[2], Tolulope Sajobi[3], Anne Gadermann[4], Juxin Liu[5], Bruno Zumbo[4], Lisa Lix[6]

[1]Trinity Western University & Centre for Health Evaluation and Outcome Sciences,
[2]Centre for Health Evaluation and Outcome Sciences, [3]University of Calgary,
[4]University of British Columbia, [5]University of Saskatchewan, [6]University of Manitoba

Sample

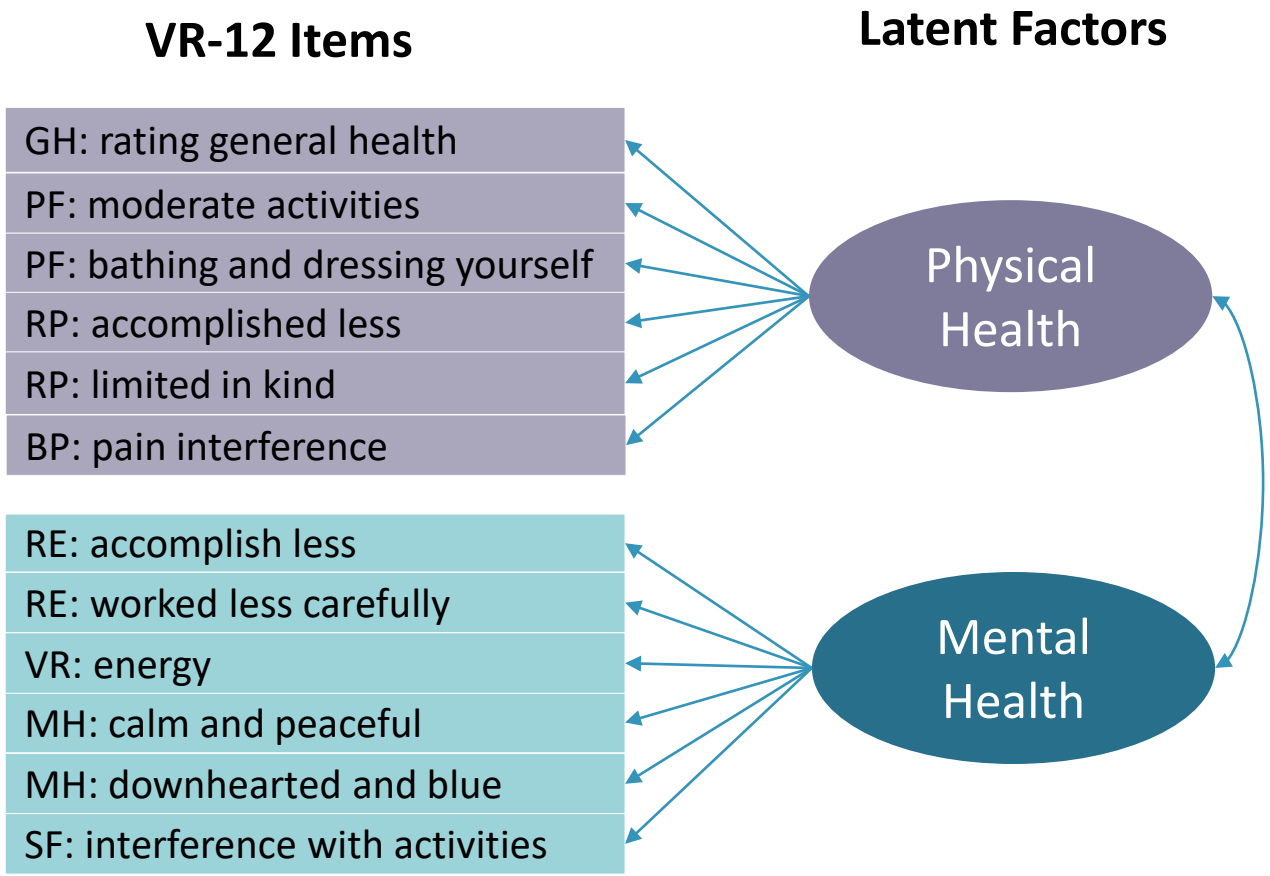
Random sample (N = 10,000) from the 2012-2014 cohort (n=296,320) of the US Medicare Health Outcomes Survey

Variables	%
Age	
<65	14.9
65-74	51.0
75+	34.0
Race	
White	80.9
African American	12.0
Other	7.1
Gender (female)	42.5
High blood pressure	33.5
Cancer history	14.6
Stroke history	9.1
Health comparison*	69.1

*good/very good/excellent

Analytical approach

Baseline measurement model

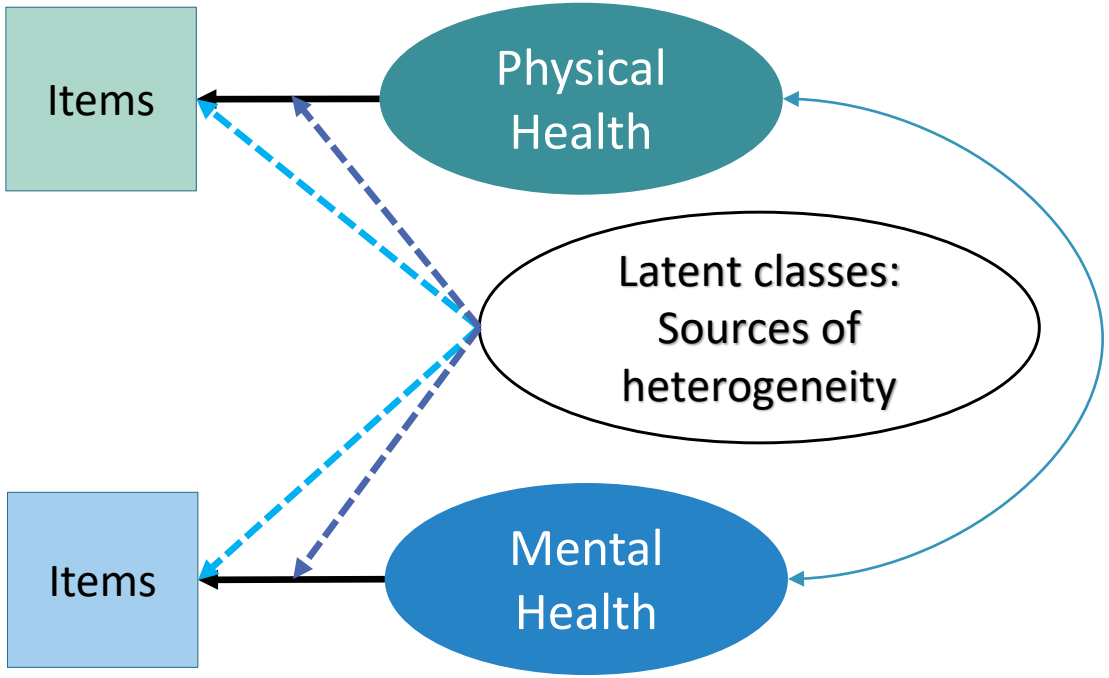


MULTIDIMENSIONAL GRADED RESPONSE MODEL

Analytical approach

Examine heterogeneity using latent variable mixture models (LVMMs)

Discrimination + difficulty parameters vary across latent classes that represent heterogeneity in the population.



MIXTURE OF THE GRADED RESPONSE MODEL

Analytical approach

Identify sources of heterogeneity

Explain latent class membership

- Logistic regression using the 3-step approach in MPlus software

Explanatory variables

- **DEMOGRAPHIC VARIABLES**:
  Age, sex, marital status, race

- **HEALTH-RELATED VARIABLES**:
  Cancer Hx, Stroke Hx, Falls, Shortness of breath, Back pain, Urinary incontinence, Difficulty reading, Difficulties with activities of daily living, Mental health compared to one year ago, Physical health compared to one year ago, Health compared to others

Results

Examine heterogeneity

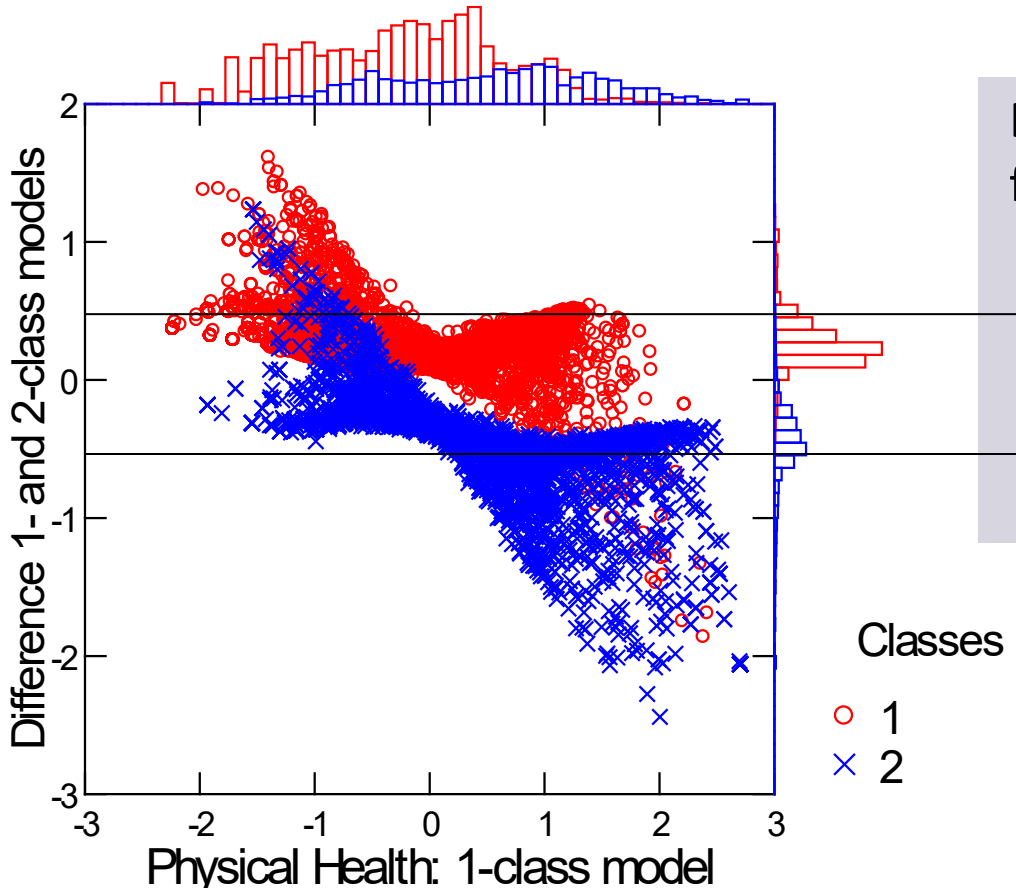
Model fit

Model	*P*	BIC	Entropy	Class proportions	
				Class 1	Class 2
1 class	60	242334			
2 classes	121	234750	.68	.66	.34

- Relative to a 1-class model (assuming no heterogeneity), improved model fit was obtained when 2 classes were specified.
- The sample is heterogeneous with respect to the original VR-12 measurement structure.

Results

Evaluate bias in physical health scores



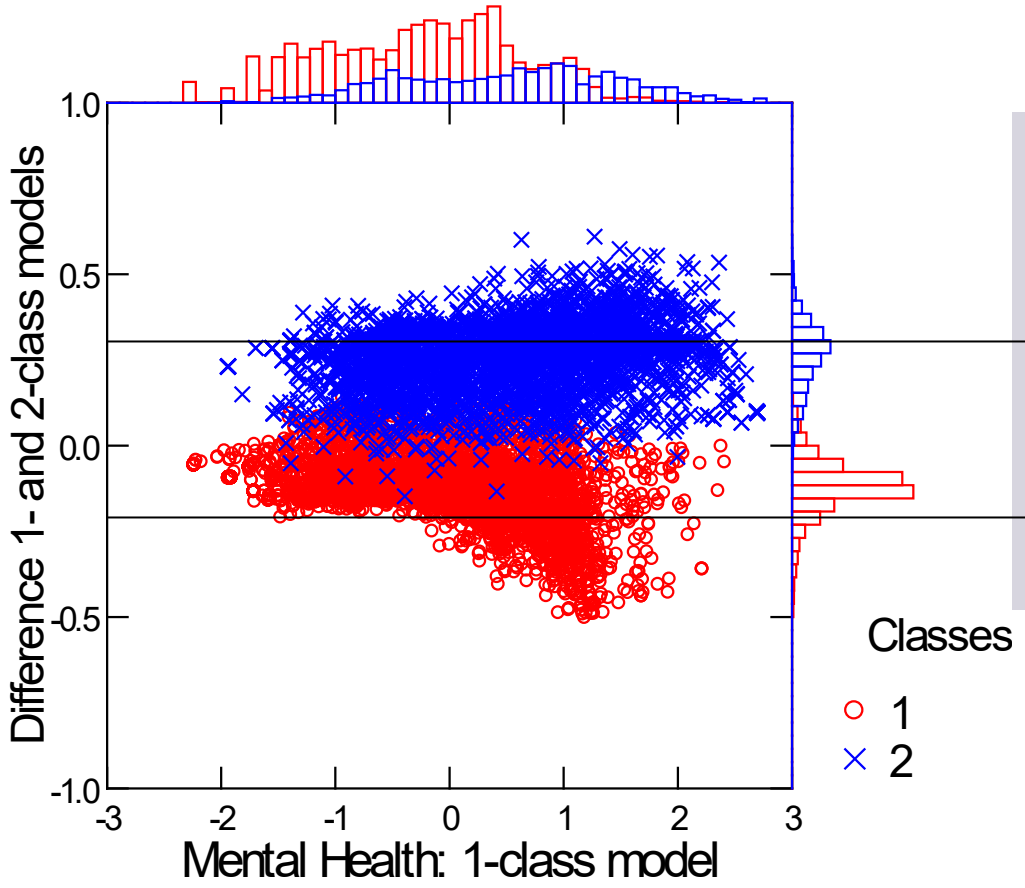
Difference in standardized factor scores of 1- and 2-class models:

10% >=.43

10% <=-.55

Results

Evaluate bias in mental health scores



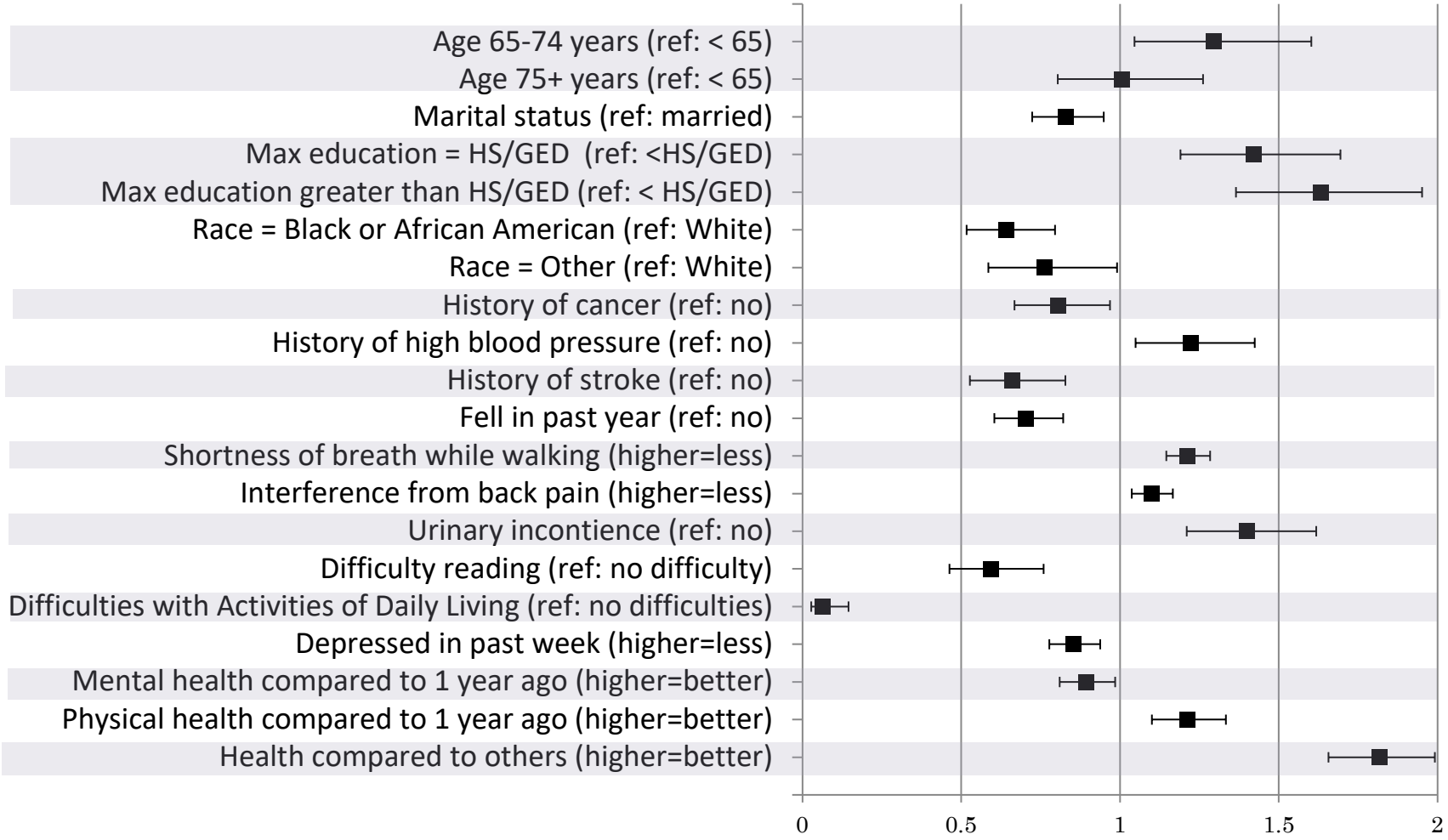
Difference in standardized factor scores of 1- and 2-class models:

10% >=.31

10% <=-.19

Results

Identify sources of heterogeneity



Conclusions

Limitations

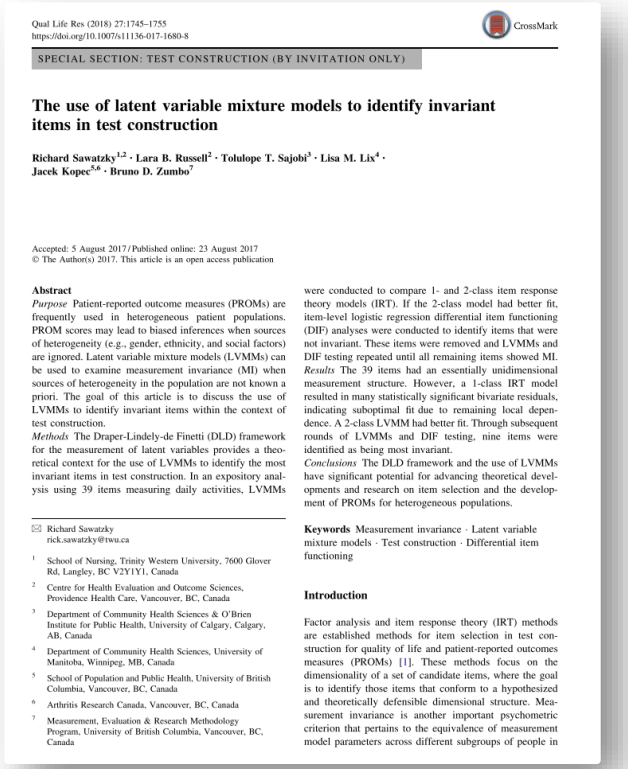
- Analyses were based on 2-class models. Further improvement may be obtained when specifying additional classes.

Ongoing research

- Models with more than 2 latent classes are being examined
- Different baseline measurement models are being compared
- Simulations studies are being conducted to evaluate the use of LVMMs for examining population heterogeneity

Use of LVMMs to identify invariant items

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction [journal article]. *Quality of Life Research, 27(7), 1745-1755.* https://doi.org/10.1007/s11136-017-1680-8



Qual Life Res (2018) 27:1745–1755
https://doi.org/10.1007/s11136-017-1680-8

SPECIAL SECTION: TEST CONSTRUCTION (BY INVITATION ONLY)

The use of latent variable mixture models to identify invariant items in test construction

Richard Sawatzky[1,2] · Lara B. Russell[2] · Tolulope T. Sajobi[3] · Lisa M. Lix[4] · Jacek Kopec[5,6] · Bruno D. Zumbo[7]

Accepted: 5 August 2017 / Published online: 23 August 2017

Abstract
Purpose Patient-reported outcome measures (PROMs) are frequently used in heterogeneous patient populations. PROM scores may lead to biased inferences when sources of heterogeneity (e.g., gender, ethnicity, and social factors) are ignored. Latent variable mixture models (LVMMs) can be used to examine measurement invariance (MI) when sources of heterogeneity in the population are not known a priori. The goal of this article is to discuss the use of LVMMs to identify invariant items within the context of test construction.
Methods The Draper-Lindely-de Finetti (DLD) framework for the measurement of latent variables provides a theoretical context for the use of LVMMs to identify the most invariant items in test construction. In an expository analysis using 39 items measuring daily activities, LVMMs were conducted to compare 1- and 2-class item response theory models (IRT). If the 2-class model had better fit, item-level logistic regression differential item functioning (DIF) analyses were conducted to identify items that were not invariant. These items were removed and LVMMs and DIF testing repeated until all remaining items showed MI.
Results The 39 items had an essentially unidimensional measurement structure. However, a 1-class IRT model resulted in many statistically significant bivariate residuals, indicating suboptimal fit due to remaining local dependence. A 2-class LVMM had better fit. Through subsequent rounds of LVMMs and DIF testing, nine items were identified as being most invariant.
Conclusions The DLD framework and the use of LVMMs have significant potential for advancing theoretical developments and research on item selection and the development of PROMs for heterogeneous populations.

✉ Richard Sawatzky
rick.sawatzky@twu.ca

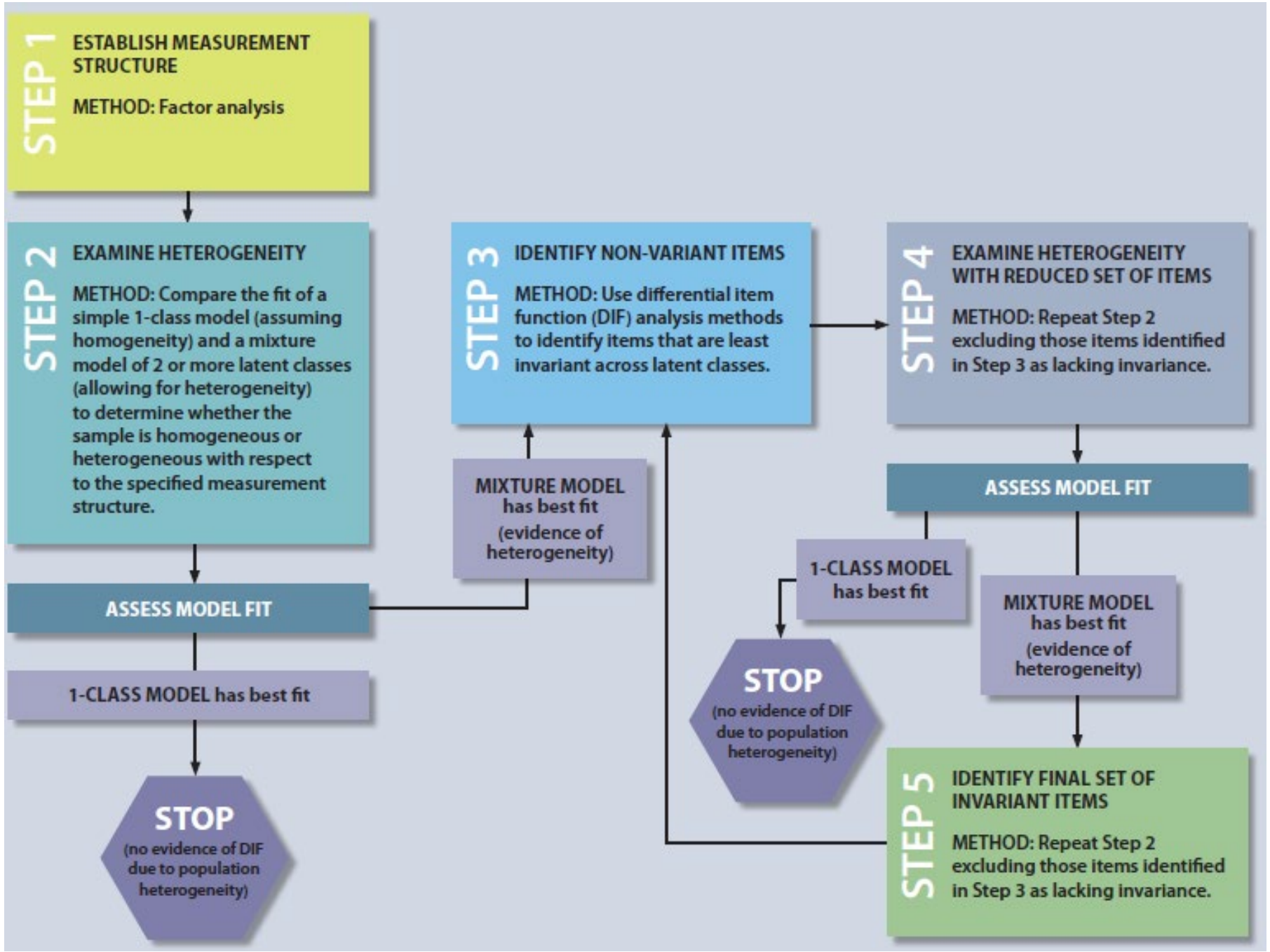
[1] School of Nursing, Trinity Western University, 7600 Glover Rd, Langley, BC V2Y1Y1, Canada
[2] Centre for Health Evaluation and Outcome Sciences, Providence Health Care, Vancouver, BC, Canada
[3] Department of Community Health Sciences & O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada
[4] Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada
[5] School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada
[6] Arthritis Research Canada, Vancouver, BC, Canada
[7] Measurement, Evaluation & Research Methodology Program, University of British Columbia, Vancouver, BC, Canada

Keywords Measurement invariance · Latent variable mixture models · Test construction · Differential item functioning

Introduction

Factor analysis and item response theory (IRT) methods are established methods for item selection in test construction for quality of life and patient-reported outcomes measures (PROMs) [1]. These methods focus on the dimensionality of a set of candidate items, where the goal is to identify those items that conform to a hypothesized and theoretically defensible dimensional structure. Measurement invariance is another important psychometric criterion that pertains to the equivalence of measurement model parameters across different subgroups of people in

APPLICATION EXAMPLES


STEP 1
ESTABLISH MEASUREMENT STRUCTURE
METHOD: Factor analysis
STEP 2
EXAMINE HETEROGENEITY
METHOD: Compare the fit of a simple 1-class model (assuming homogeneity) and a mixture model of 2 or more latent classes (allowing for heterogeneity) to determine whether the sample is homogeneous or heterogeneous with respect to the specified measurement structure.
ASSESS MODEL FIT
1-CLASS MODEL has best fit
STOP
(no evidence of DIF due to population heterogeneity)
MIXTURE MODEL has best fit
(evidence of heterogeneity)
STEP 3
IDENTIFY NON-VARIANT ITEMS
METHOD: Use differential item function (DIF) analysis methods to identify items that are least invariant across latent classes.
STEP 4
EXAMINE HETEROGENEITY WITH REDUCED SET OF ITEMS
METHOD: Repeat Step 2 excluding those items identified in Step 3 as lacking invariance.
ASSESS MODEL FIT
1-CLASS MODEL has best fit
STOP
(no evidence of DIF due to population heterogeneity)
MIXTURE MODEL has best fit
(evidence of heterogeneity)
STEP 5
IDENTIFY FINAL SET OF INVARIANT ITEMS
METHOD: Repeat Step 2 excluding those items identified in Step 3 as lacking invariance.

Approach

Methods

Sample

1,666 adults living in British Columbia, Canada.

- Mean age = 56.7 years (SD=15.9)
- 60.7% female
- 28.3% treated for rheumatoid arthritis
- 38.2% treated for osteoarthritis
- 74.0% had another health condition
- 24.0% reported “fair” or “poor” health

Measure

Daily Activities item bank (39 items) of the CAT-5D-QOL (Kopec et al. 2006)

Measures limitations and need for assistance in ability to perform usual activities, using 3-, 4- and 5-point Likert type response scales.

Results

1	Exploratory (EFA) and confirmatory (CFA) factor analyses or Item Response Theory (IRT) to establish unidimensionality.	**Essential uni-dimensionality was supported:** • EFA: Ratio of 1st to 2nd eigenvalue was 16. • CFA: Acceptable fit of a unidimensional model (RMSEA = 0.082 and the CFI = 0.986).
2	Fit of an IRT Graded Response Model (GRM) in the overall sample. LVMM: Discrimination and difficulty parameters of the GRM were allowed to vary across 2 latent classes. Examining heterogeneity: Information criteria and indicators of fit were used to compare 1- and 2-class models.	**Fit of 2-class LVMM was better than the 1-class GRM:** • Bayesian Information Criterion (BIC) 2-class LVMM (75287) < than BIC 1-class GRM (77000). • Statistically significant ($p<.05$) bivariate residuals were reduced from 45.5% of all item pairs in the 1-class GRM to 9.5% in the 2-class LVMM. • Class proportions: 0.64 and 0.36

Results

3	Ordinal logistic regression (OLR) approach was used to identify items that exhibit DIF. Each item was regressed on: (i) the latent factor score and (ii) the latent factor score plus latent class membership and their interaction. Least invariant items were identified based on a difference in explained variance (ΔR2) greater than 0.035 when comparing (i) and (ii).	**23 items were identified as least invariant:** • OLR identified 23 items with ΔR2 > 0.035. The resulting 16-item model was retested in Step 4.
4	The 1-class GRM and 2-class LVMM were re-estimated excluding the least invariant items from Step 3.	**Fit of 16-item 2-class LVMM was better than the 1-class GRM:** • BIC 2-class LVMM (26781) < than BIC 1-class GRM (27268). • % statistically significant bivariate residuals was reduces from 30.5% for the 1-class GRM to 5.0% for the 2-class LVMM. • Class proportions: 0.74 and 0.26

Results

5	The process was stopped when none of the items exceeded the ΔR2 cut-off in Step 3.	**5 of the 16 items were identifies as least invariant:** • OLR (next iteration of Step 3) identified 5 items with ΔR2 > 0.035. • Step 4 was applied to the remaining 9 items. The sample was homogeneous with respect to the 9 items • Fit of 9-item 2-class LVMM was worse than the 1-class GRM: BIC 2-class LVMM (13942) > than BIC 1-class GRM (13862).

Conclusions

- 9 invariant items
- Need for replication in other diverse populations
- Other considerations pertaining to item content and construct validity need to be considered.

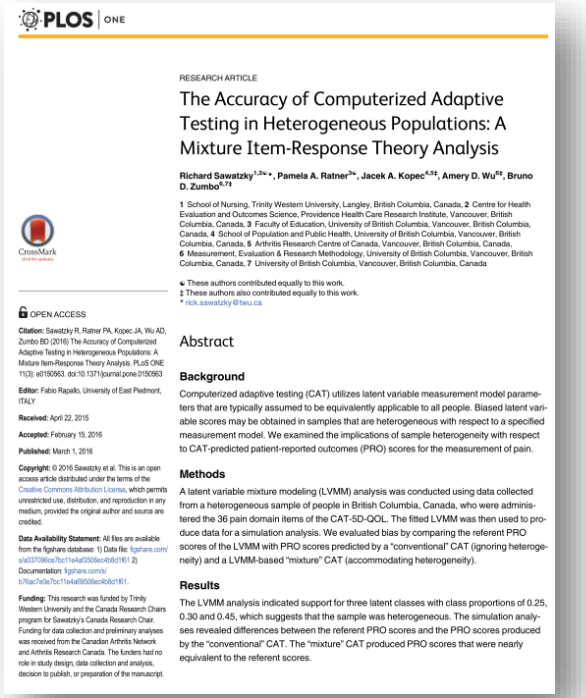
Item content (# response options)

- Difficulty preparing one's own meals (5)
- Difficulty grooming oneself (5)
- Difficulty performing light household chores (5)
- Difficulty bathing oneself without help (5)
- Difficulty dressing and undressing oneself (5)
- Need for help with getting around the house (5)
- Ability to take care of oneself (3)
- Difficulty using the toilet (5)
- Need for help with using the toilet (5)

Use of LVMMs to adjust scoring + selection of items in computerized adaptive tests

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.



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

Citation: Sawatzky R, Ratner PA, Kopec JA, Wu AD, Zumbo BD (2016) The Accuracy of Computerized Adaptive Testing in Heterogeneous Populations: A Mixture Item-Response Theory Analysis. PLoS ONE 11(3): e0150563. doi:10.1371/journal.pone.0150563

Editor: Fabio Rapallo, University of East Piedmont, ITALY

Received: April 22, 2015

Accepted: February 15, 2016

Published: March 1, 2016

Data Availability Statement: All files are available from the figshare database: 1) Data file: figshare.com/s/a03709bce7bc11e4af3506ec4b8d1f61 2) Documentation: figshare.com/s/b76ac7e0e7bc11e4af3506ec4b8d1f61.

Funding: This research was funded by Trinity Western University and the Canada Research Chairs program for Sawatzky's Canada Research Chair. Funding for data collection and preliminary analyses was received from the Canadian Arthritis Network and Arthritis Research Canada. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

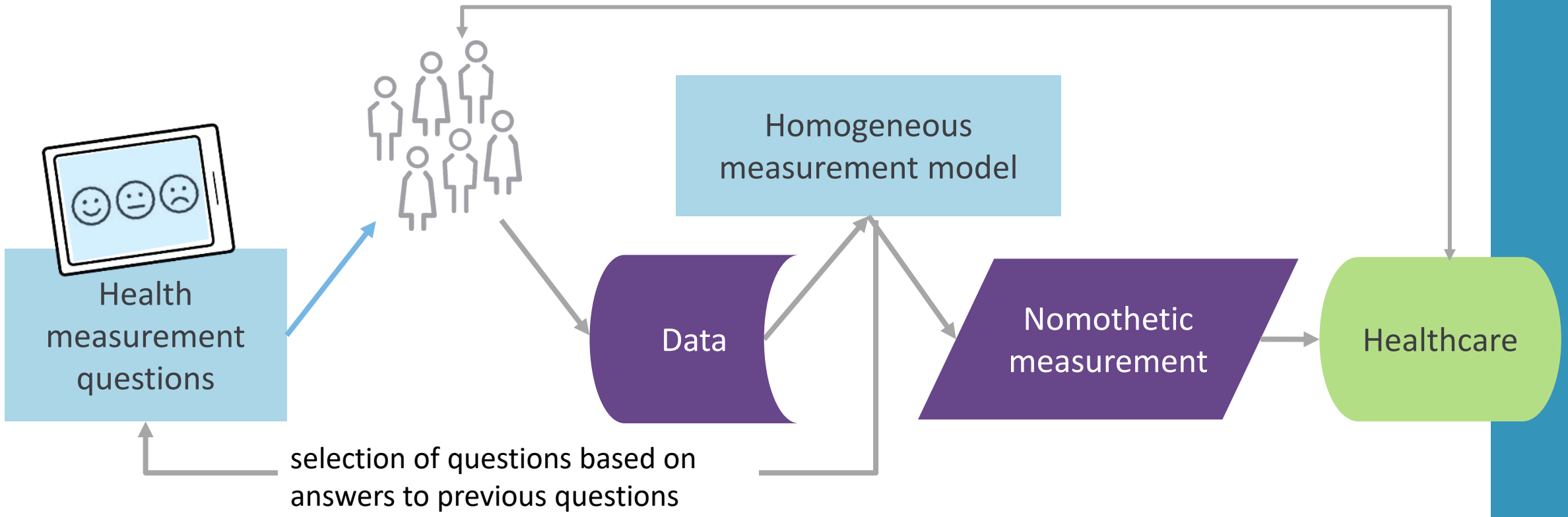
The LVMM analysis indicated support for three latent classes with class proportions of 0.25, 0.30 and 0.45, which suggests that the sample was heterogeneous. The simulation analyses revealed differences between the referent PRO scores and the PRO scores produced by the "conventional" CAT. The "mixture"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

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.* https://doi.org/10.1371/journal.pone.0150563

APPLICATION EXAMPLES

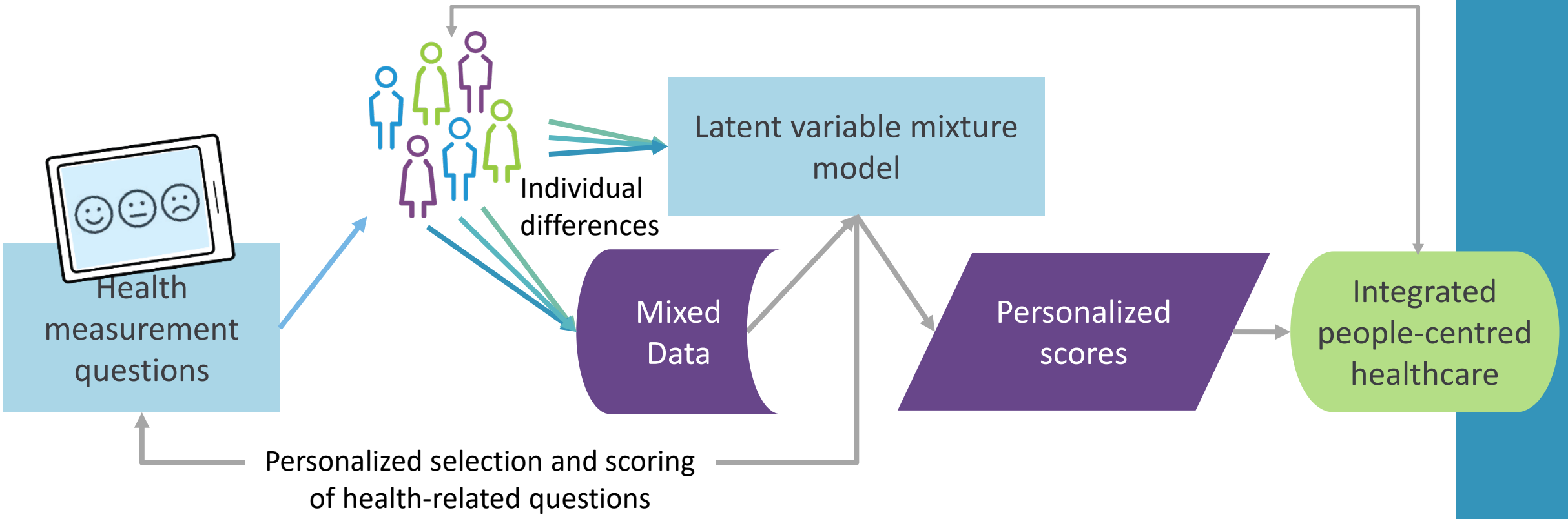
Conventional “nomothetic” health measurement CAT

‘one size fits all’



Equitable people-centred health measurement CAT

'one size does not fit all'



“Mixture” measurement model

‘one size does not fit all’

- The cumulative probability of an item response at or above category *j* within a latent class can be computed as follows:

$$P_{ijk}\left(Y \geq j \mid \theta, C = k\right) = \frac{\exp(-\tau_{ijk} + \lambda_{ik}\theta)}{1+\exp(-\tau_{ijk} + \lambda_{ik}\theta)}.$$

- Each class has a unique set of parameters that are estimated simultaneously in the latent variable mixture model:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

, where *f* is the mixture of the class-specific distributions, and π_k is the mixing proportion.

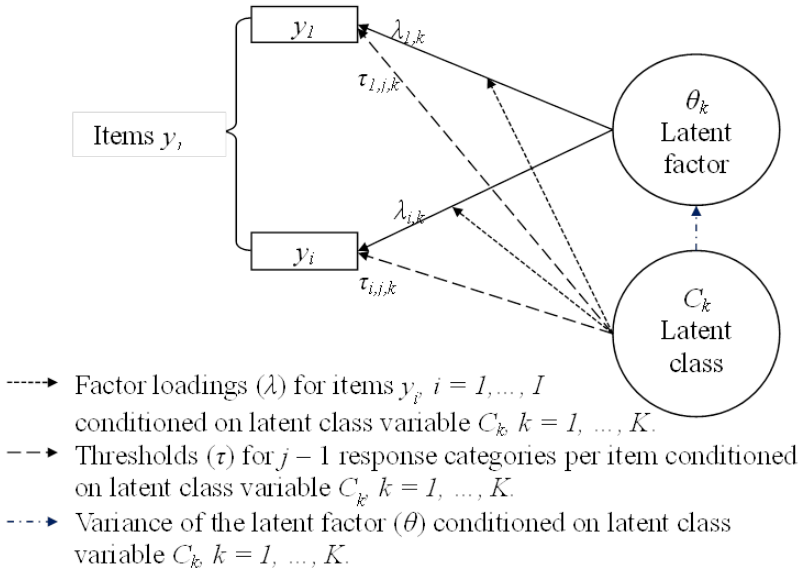
- The cumulative probability of an item response at or above category *j* within a latent class can be computed as follows:

$$P_{ijk}\left(Y \geq j \mid \theta, C = k\right) = \frac{\exp(-\tau_{ijk} + \lambda_{ik}\theta)}{1+\exp(-\tau_{ijk} + \lambda_{ik}\theta)}.$$

- The cumulative probability of an item response at or above category *j* within a heterogeneous population is obtained by:

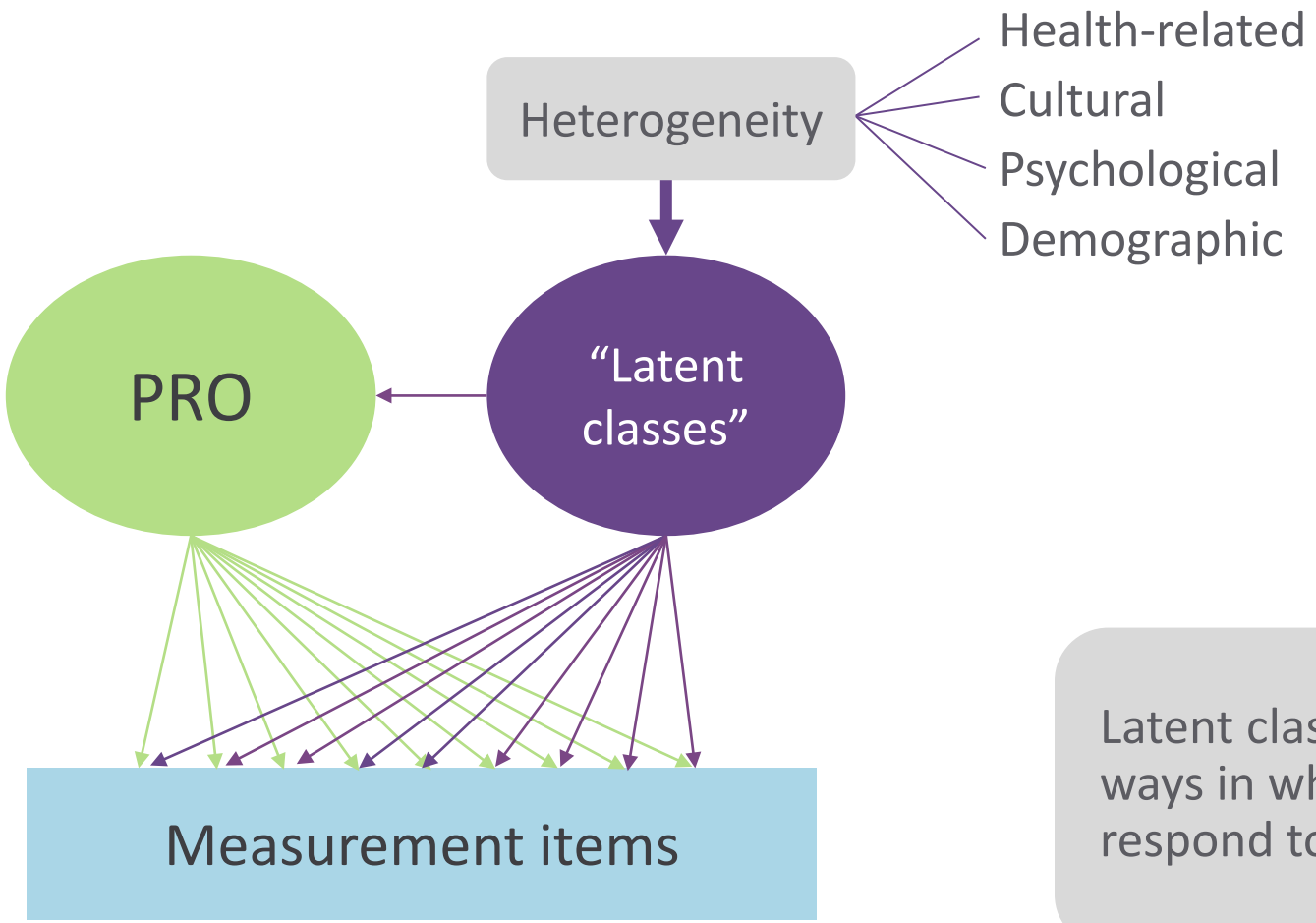
$$P_{ij}(Y \geq j \mid \theta) = \sum_{k=1}^{K} (X_k * P_{ijk}(Y \geq j \mid \theta))$$

, where X_k is the posterior probability of an individual being in class *k*.



“Mixture” measurement model

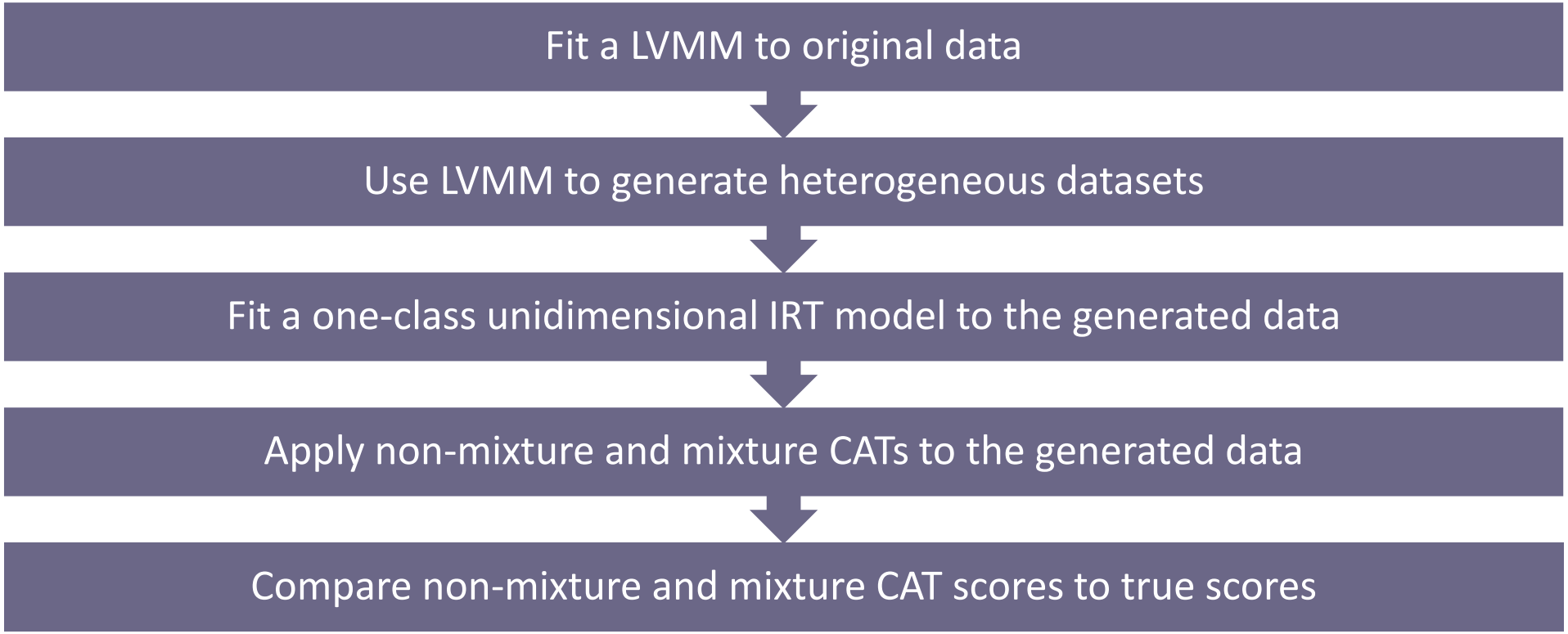
‘one size does not fit all’



Latent classes represent different ways in which people interpret and respond to questions

Methods

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



Fit a LVMM to original data

Item bank measuring pain or discomfort

- 36 items measuring the severity and frequency of pain or discomfort & the impact of pain on activities of daily living and leisure
- One of the item banks of the CAT-5D-QOL (Kopec et al., 2006)

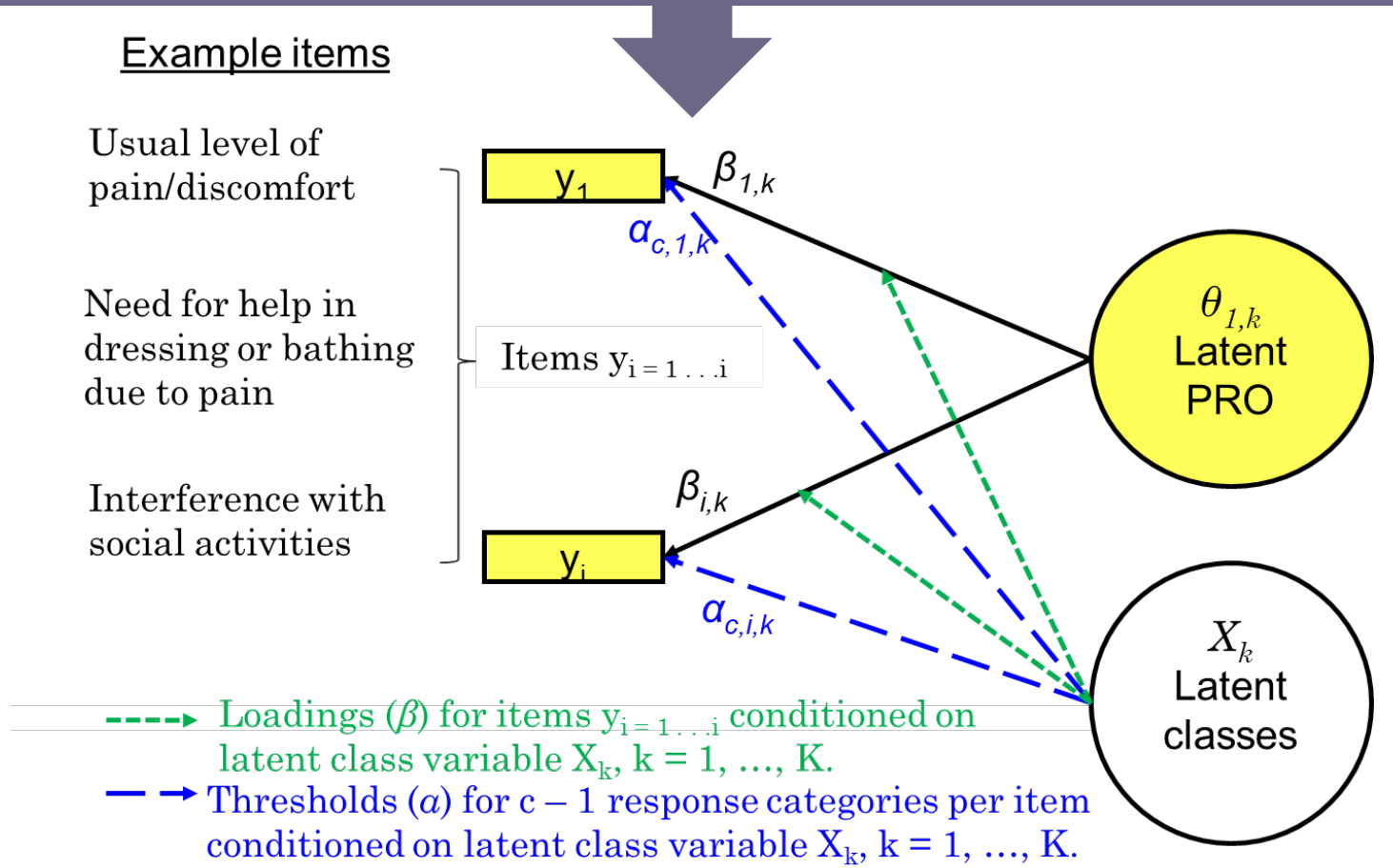
Sample

- 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



Generate heterogeneous data

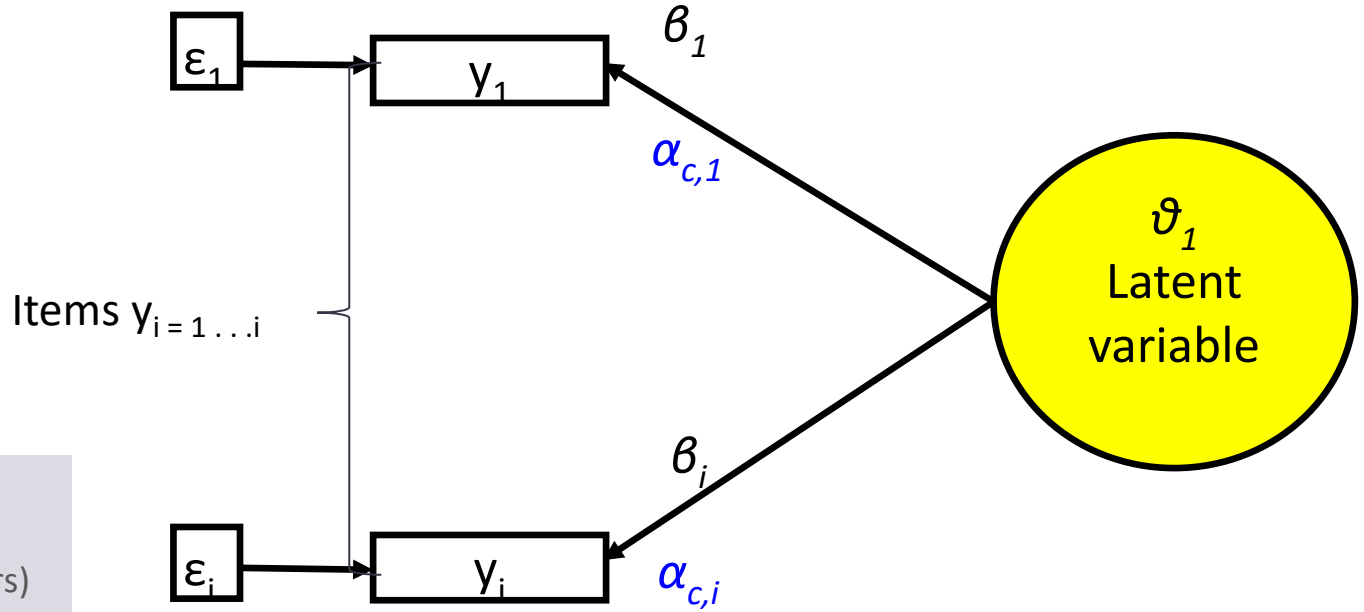
100 datasets (of N=1,000 each) were generated using

- Parameters based on the LVMM
- Randomly sampled PRO scores (“true scores”) based on the distribution of LVMM-predicted 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	Class 3
1 class	177	87884					
2 classes	354	86056	3141	0.86	0.59	0.41	
3 classes	531	85654	1713	0.83	0.25	0.30	0.43

N = 1,662. *P* = number of model parameters. BIC = Bayesian Information Criterion. LL = log likelihood
[1] Likelihood ratio of k and k-1 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 3 classes were specified.
- The sample is not homogeneous with respect to the pain item bank.

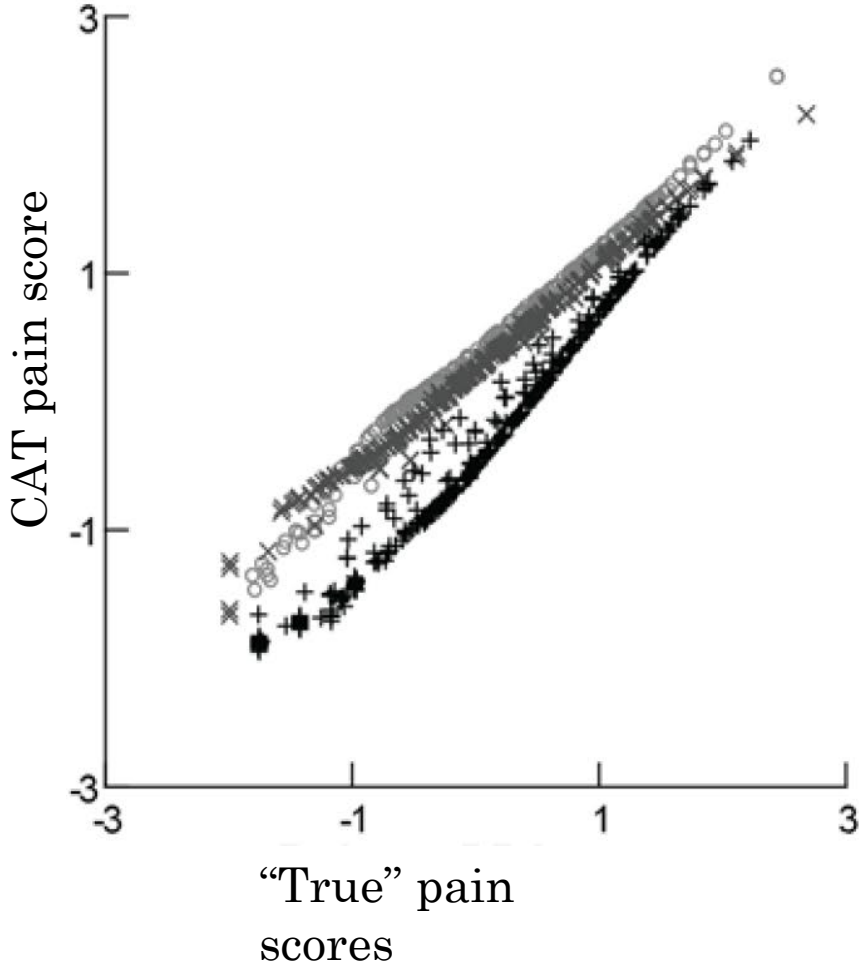
Description of latent classes

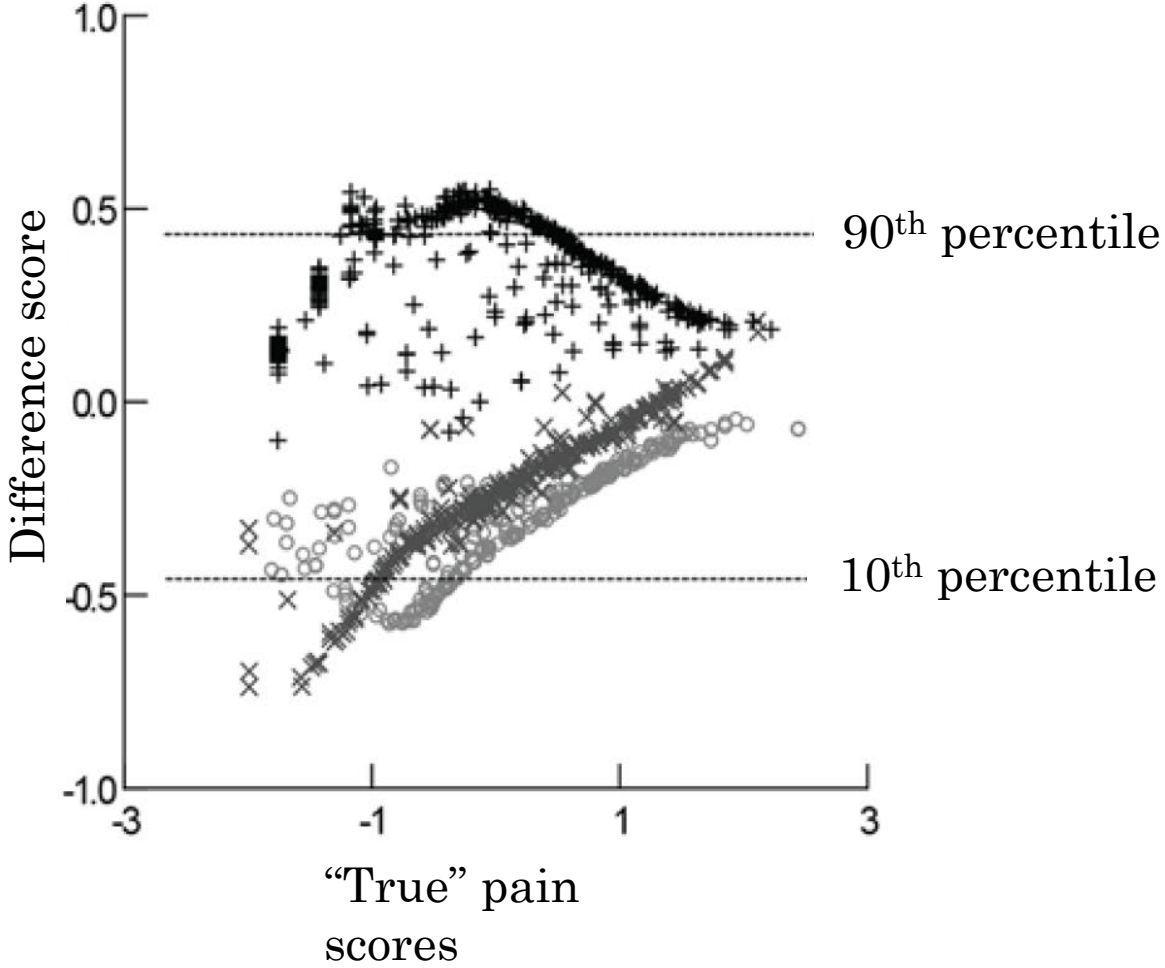
Variables	Prevalence				Multivariate logistic regression[b]		
	Full sample	Class 1[a]	Class 2[a]	Class 3[a]	OR (95% CI) classes 1 versus 3	OR (95% CI) classes 2 versus 3	OR (95% CI) classes 2 versus 1
Sex (referent = male)	60.6	63.3	63.2	57.1	1.1(0.8;1.4)	1.0(0.8;1.5)	1.1(0.8;1.4)
Age (mean (sd))[c]	57(15.9)	58.3 (17.5)	56.9 (16.0)	55.3 (17.3)	1.0(0.9;1.1)	0.9(0.8;1.0)	0.9(0.8;1.0)
Taking medications	77.9	85.8	84.2	67.9	1.8(1.2;2.7)	1.8(1.3;2.7)	1.0(0.6;1.7)
Hospitalized during past year	20.5	27.2	19.2	17.3	1.2(0.9;1.7)	0.8(0.6;1.2)	0.7(0.5;1.0)
Has rheumatoid arthritis	28.0	37.4	27.9	21.9	1.2(0.7;2.0)	1.4(0.9;2.3)	1.2(0.7;2.1)
Has osteoarthritis	36.6	40.7	45.5	27.9	1.5(1.0;2.0)	2.1(1.4;3.0)	1.4(1.0;2.1)
Has another health condition	77.3	81.1	83.6	70.6	1.3(0.9;1.8)	1.5(1.1;2.2)	1.2(0.8;1.9)
Self-reported health is fair or poor (referent = good, very good or excellent)	24.0	32.7	27.0	16.5	1.6(1.1;2.3)	1.5(1.1;2.1)	0.9(0.7;1.3)
Sampling groups							
Community-dwelling (referent)	59.8	48.7	55.8	67.5	1.0	1.0	1.0
Rheumatology clinic sample	20.4	29.7	16.7	15.7	1.4(0.8;2.5)	0.6(0.3;1.0)	0.4(0.2;0.7)
Awaiting joint replacement surgery sample	19.8	21.6	24.5	16.8	0.9(0.6;1.5)	0.8(0.5;1.2)	0.8(0.5;1.4)

Notes. OR = odds ratio. N = 1,660 a Prevalence computed based on posterior-probability based multiple imputations using the Mplus software. Proportions of latent class membership are .27, 30 and .43 for classes 1, 2 and 3, respectively. b Odds ratios based on the multinomial logistic regression using pseudo-class draws. c For each 10-year (decade) increase in age.

Description of latent classes

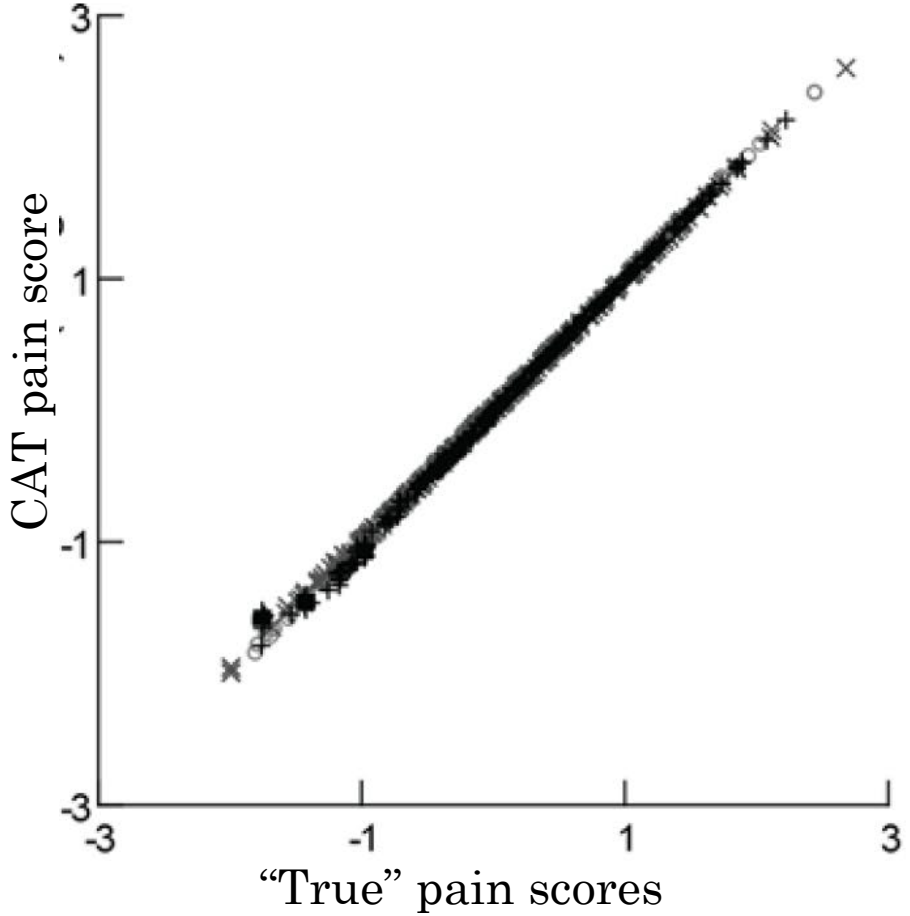
Ignoring heterogeneity (based on 1-class GRM)

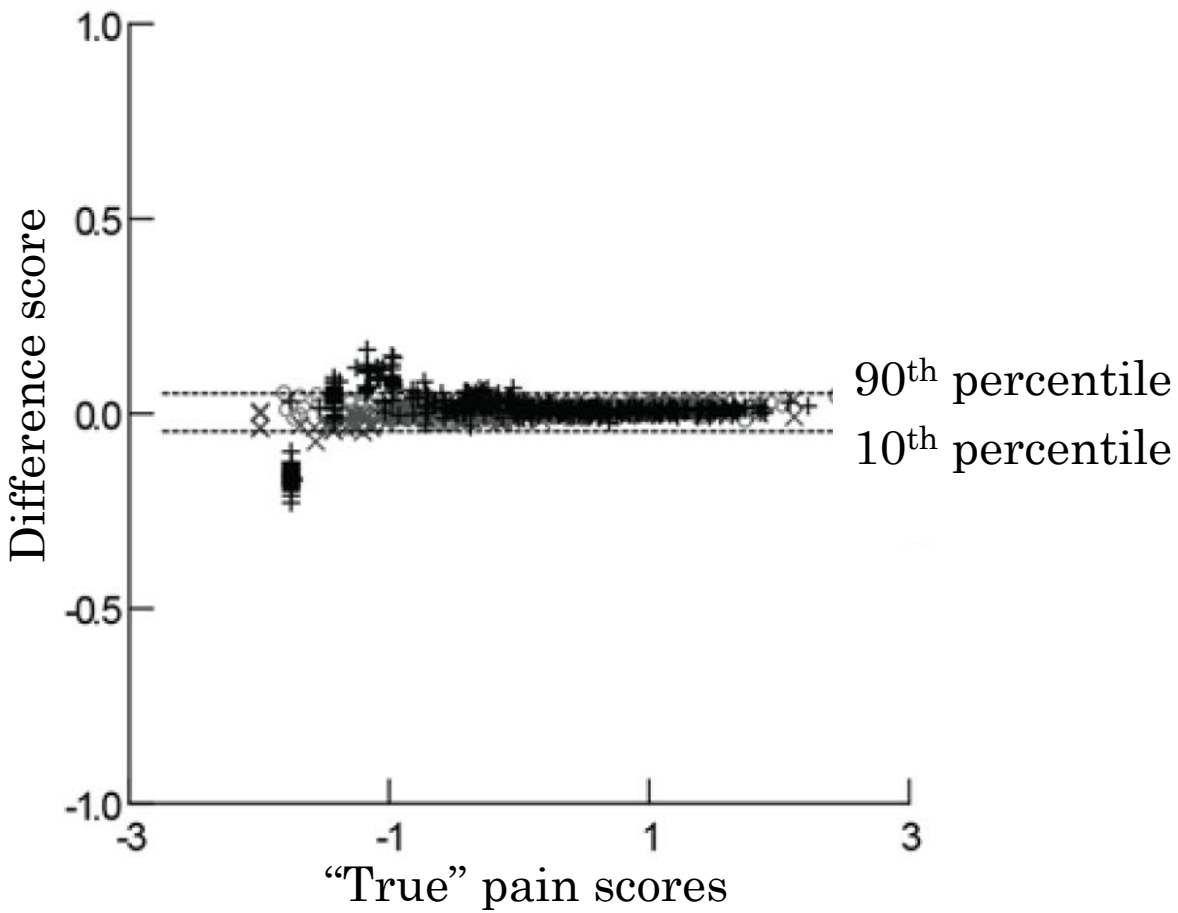




Mixture CAT-predicted scores

Adjusting for heterogeneity (based on 3-class LVMM)





Discussion

The challenge of population heterogeneity

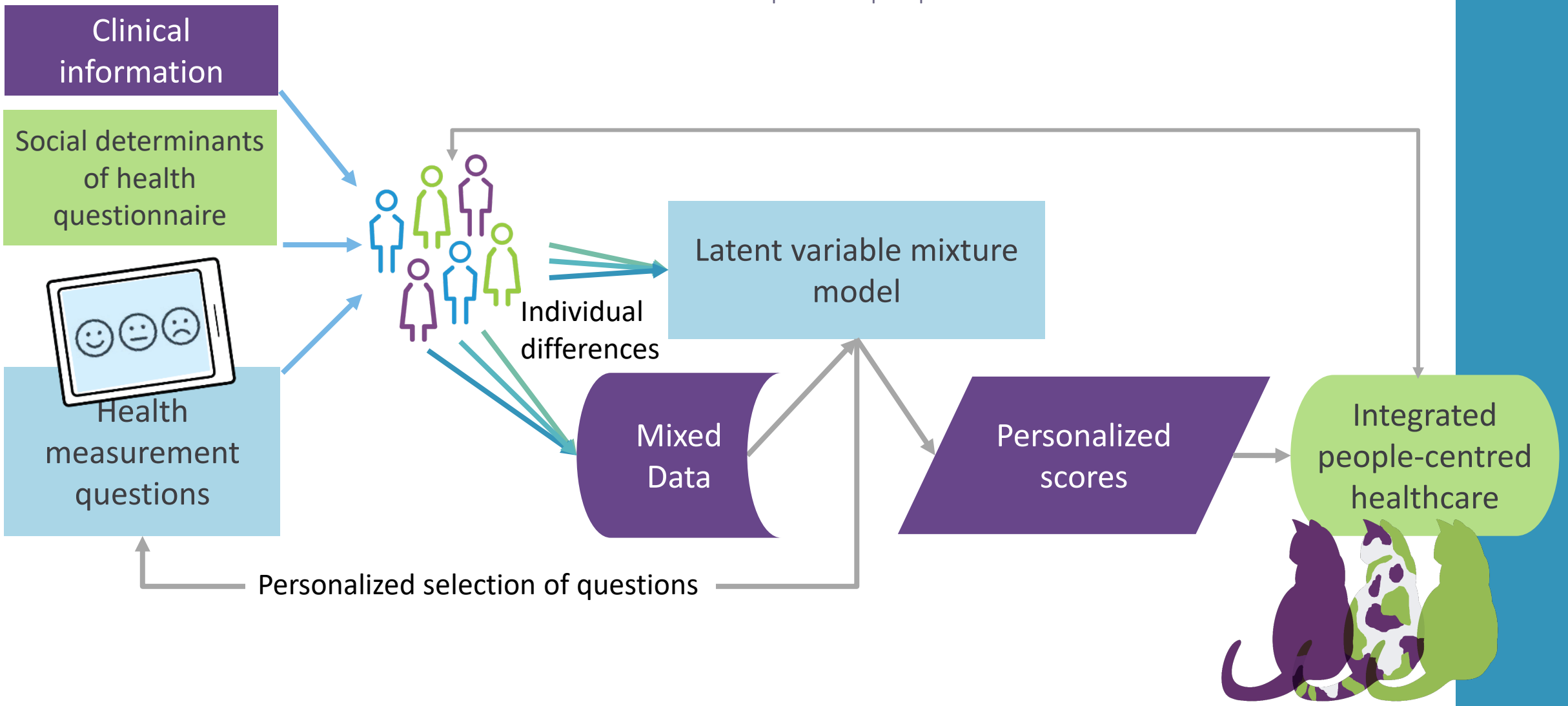
- People may not interpret and respond to questions about their health and quality of life in the same way.
- Unaccounted for heterogeneity could result in measurement biases that could negatively affect some people.

Further research

- Need to account for individual differences in health measurement
- Person-centred, versus variable-centred, statistical methods offer a potential solution by using LVMMs to measure and analyze PROs

Current research on mixture-CAT

Methods for equitable people-centred health measurement



Current research on mixture-CAT

Methods for equitable people-centred health measurement

Aim: To further develop and evaluate our equitable people-centred health measurement (EPHM) methodology that tailors PRO measurements by integrating information about SDOH in diverse populations into a “mixture-CAT”.

Develop

- Further develop mixture-CAT for measuring pain and emotional wellbeing by integrating SDOH and health-related predictors of measurement bias.

Evaluate

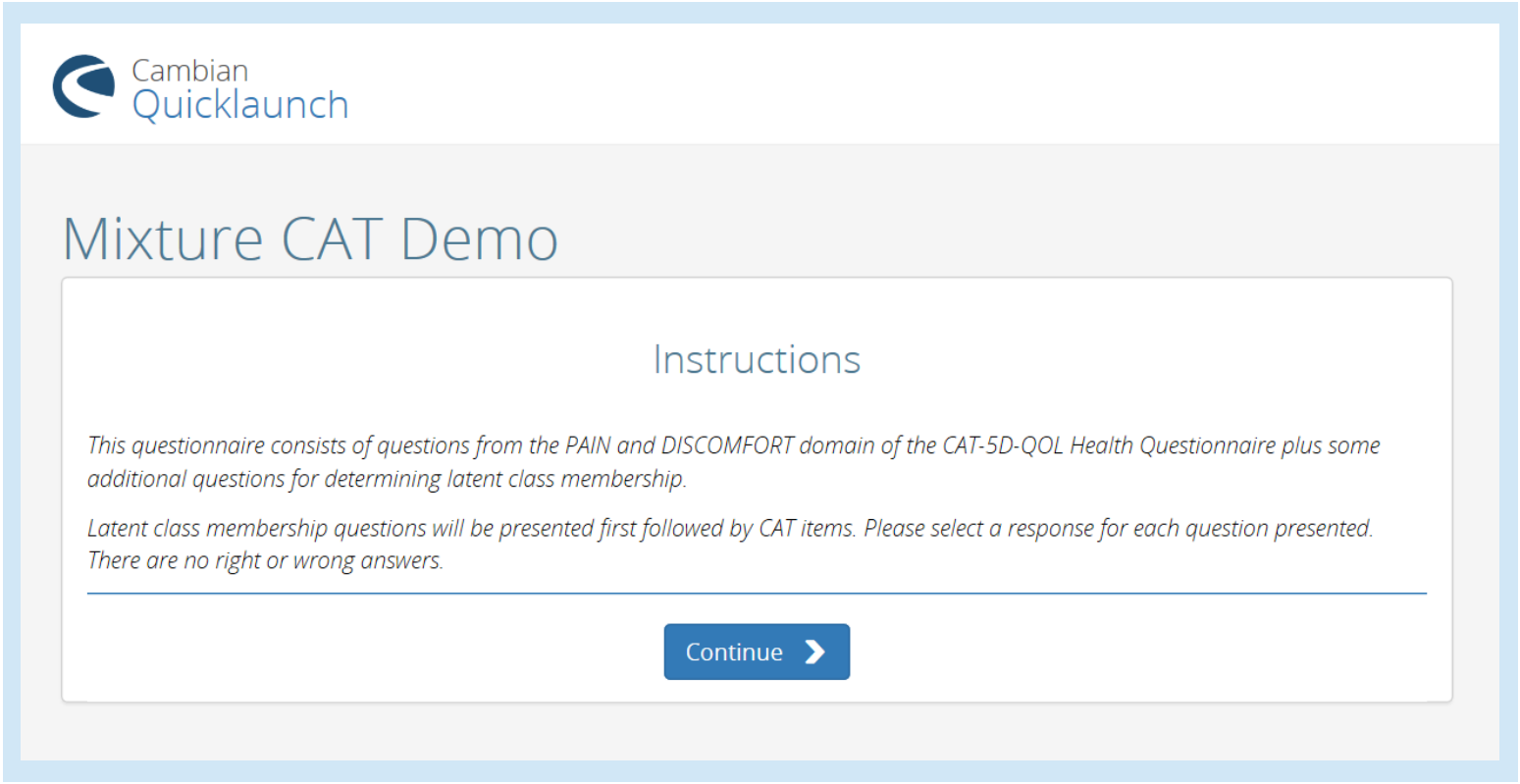
- Evaluate mixture-CAT by comparing the predictive performance of PRO trajectories over 6 months across 4 modalities: 1) mixture-CAT, 2) CAT, 3) fixed-form PROMs, and 4) single-item PROMs.

Implement

- Determine knowledge users’ perspectives about the use of mixture-CAT results using an integrated knowledge translation approach

Mixture-CAT prototype

Demo



https://www.cambian.com/mixture-cat-demo/

THANK YOU!

Richard Sawatzky, RN, PhD

Professor & Canada Research Chair in Person-Centred Outcomes
School of Nursing, Trinity Western University, Langley BC

Centre for Health Evaluation and Outcomes Sciences,
St. Paul's Hospital, Vancouver, BC

Salgrenska Academy, Centre for Person-Centred Care,
University of Gothenburg, Sweden

rick.sawatzky@twu.ca

CHÉOS
Centre for Health Evaluation & Outcome Sciences

TRINITY WESTERN UNIVERSITY