Longitudinal Rasch Model in Harmonizing Adherence Recall Intervals

Yan Wang* Honghu Liu†

Abstract

The data harmonization aim to create the common measure that can be used across studies. Harmonization is to combining data sets collected at different times into a single, consistent data series. Rasch model is not sensitive to the sample that generate the response conversion key. Recent literature conceptualize longitudinal Rasch model as the hierarchical generalized linear models by assuming the latent variable as random effect. In longitudinal data harmonization, the challenge part is to calibrate the latent trait over time, estimated the invariant parameters in the Rasch model and preserve the changing of the trait for modeling. Self-reported adherence is the most convenient way to obtain the adherence measure among HIV patients by directly asking the patients themselves. In this study, we use the longitudinal version of Rasch model to create the conversion key, which is used to calibrate the self-reported adherence of HIV medication over time. The results are compared with the device measured adherence measure to ensure the validity of calibration.

Key Words: Harmonization, Longitudinal, Self-reported, Item Response Theory, Rasch Model, Adherence

1. Introduction

1.1 Background

Item Response Theory (IRT) is a big class of methods that can be used to harmonize measures together with the key assumption of the latent variable. It was first used in education testing to relate the probability of correct answer to a question and the potential ability. IRT models are used to describe the probability relationship between responses to survey items and continuous latent variable [1, 2]. Later, the IRT models were extended to the field of physical functioning and psychological well-being, as well as in health literature. It was used to describe the relationship in probability between a set of items and the continuous latent variable. It was originally developed within educational research and later extended to field of physical functioning and psychological well-being. In recent decades, more and more IRT models were applied to health research, such as relate patient reported outcomes (PRO) to quality of life in the PROMIS (patient report outcome measurement information system).

Polytomous items are more commonly used than dichotomized outcomes with only true or false scenario. This is more common for health outcomes with five likert scales, such as strongly agree, somewhat agree, neutral, somewhat disagree, and strongly disagree. The extension version of Rasch model can be used to handle the polytomous situation.

IRT models has two unique and important advantages when harmonize measures from different items [3]. The first is to create the common score across surveys but retain the additional information from survey-specific items, which will eventually enhance the score precision. The other unique property is that IRT allows different item functioning (DIF). The DIF arises because of different culture, language settings, resulting in different interpretaion of survey items. IRT will take this into account when generalizing the harmonized

*Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, CA 90095

†Division of Public Health and Community Dentistry, UCLA School of Dentistry, David Geffen School of Medicine, Fielding School of Public Health, Los Angeles, CA 90095

common scale. The summarized score methods without DIF inflate the group differences and report misleading conclusions because of the unnecessary measurement error [3].

Recently IRT models have been used increasingly in health status measurements to evaluate patient reported outcomes (PROs) [4, 5, 6, 7]. Longitudinal design is a more generally used study settings for health related studies to evaluate the change of health outcomes. Patient-reported outcomes (PROs) are used to capture patients self evaluation of a disease status and quality of life.

1.2 Adherence

Patients with chronic disease had poor adherence as low as 50% [8]. The general universal optimal cut point is between 63% and 89%, measured by medication possession ratio and proportion of days covered with prediction of any cause hospitalization [9]. In prediction of disease-specific hospitalization, the optimal cut-off adherence is from 58% to 85% [9]. For HIV adherence, the well recognized cut off is 85% [10] with the consideration of optimal clinical outcome [11]. More importantly, when adherence is less than 50%, the patients are at great risk for treatment failure and disease progression [10]. ART adherence less than 50% is associated with lower risk for treatment resistant viral mutations than adherence between 50% and 95%.

1.3 Harmonization

According to Business Dictionary, harmonization is defined as the *adjustment of differences and inconsistencies among different measurements, methods, procedures, schedules, specifications, or systems to make them uniform or mutually compatible* [12]. The definition potentially illustrates the process and development of harmonization of data, from measurements to system [13]. The basic and fundamental harmonizing process is measurements. The aim of this process is to generalize uniform measure, adjust for the *differences* and *inconsistencies*, in order that different measures for the same contents are compatible.

Data Harmonization is a topic rapidly growing from recent decades as the needs of building large volume data system from sharing existing data resources. Data Sharing for Demographic Research (DSDR) defines this process as all efforts related to combine data from different sources and provide users with a comparable view of data from different studies. The modern research of big data system requires 4 Vs of data, volume, variety, velocity, veracity [14].

"*Harmonization is a process composed of a series of complementary steps which must be applied with rigorous procedures and decision-making in order to ensure validity and reproducibility of harmonization outputs*". [15] Harmonization of data is the usual way to improve the comparability between studies. If harmonization cannot be done, then the options are restricted to either making unwarranted assumptions about the data, or not doing any comparison at all [16]. Data harmonization is limited as a method for achieving comparability between studies. Pooling data together using appropriate methods is able to generate rich and available data for new hypothesis. For example, in the Comparison of Longitudinal European Studies (CLESA), the cross-national longitudinal data on health and functioning among older people are available after harmonized the data from six studies [17]. However, different methodologies and sampling techniques used in the different studies need a series of decisional strategies for the preparation of unbiased comparison across studies [15].

2. Methods

Several methods of data harmonization have been proposed in the literature. They can be categorized into three classes [18]. The first class of methods are relative simple and straight forward. It creates a common metric for combining constructs measure using different scales. It directly operated on the original measures or items. By certain conversion formulas of monotone functions, into a comparable measure of the original variables. For example, re-categorization, Z-score transformation, and percentile conversion are all belong to this class of method [18].

The second class of methods involves using multiple imputation techniques. The imputed value or the estimated value of the missing item are used to replace the missing item. The method is efficient for up to 50% missing data, for certain group of sample [18]. The third class of methods assumes that there is a latent factor. The measures from the different data sets are measuring the same latent variable. The term latent means the true value for the item is unknown, but can be observed through some responses of the items. Ma, Raina and Griffith use a simulation study to compare the performance of these harmonization methods [18]. Item response theory (IRT) models have been used in health status measurement and evaluation of Patient-Report Outcomes (PROs). In report of the health status, the polytomous items are common. Polytomous Rasch model is commonly used in psychometric to find the latent common measure.

2.1 Rasch Model

Item response theory (IRT) was used in health to interpret the probabilistic relationship between a set of patient report outcomes (PRO) and the unidimensional latent variable. The traditional approach to summarize the PRO information had drawbacks, such as no individual level information with ordinal scores. Rasch model could overcome these issues [19].

Rasch model is the simplest IRT model among the families. We modify the available SAS macro to create the common latent trait for the self-reported adherence measures. The dichotomous version and polytomous version [20, 21] are both available for Rasch model. The Polytomous Rasch Model (PRM) using conditional maximum likelihood estimation (CMLE) is well developed [22].

This paper used polytomous longitudinal Rasch models (plRM) with marginal maximum likelihood (MML) to model the longitudinal measure - patient reported adherence.The unique assumption is at each time point of the repeated measures, basic assumption for Rasch model should be satisfied [1].

2.2 Unidimensional Rasch Model

The assumptions for simplest Rasch model can be summarized as [1] as unidimensionality, monotonicity, local independence and no differential item functioning (DIF). These assumptions has to be met before fitting the Rasch model. These tests are available for these assumptions [22, 23]. Assume $i \in I$ for item i, with $m_i + 1$ categories, denoted by $0, 1, \cdots, m_i$. The polytomous Rasch model is given by [22],

$$P(X_i = x_i|\theta) = \frac{\exp(x_i\theta + \eta_{ix_i})}{K_i} \tag{1}$$

Here K_i is,

$$K_i = \sum_{x_i=0}^{m_i} \exp(x_i\theta + \eta_{ix_i})$$

In Rasch mode, there is no discrimination parameter in the model. This means we do not discriminate the items in the model, unlike in 2PL model or graded response model. We focus on the threshold parameter, which has more emphasis on the response categories of each item.

2.3 Multidimensional Rasch Model

Multidimensional Rasch model is an extension version of unidimensional model, by plitting the set of items into $I = I_1 \cup I_2$. The items in I_1 measure θ_1 and the items in I_2 measure θ_2. We assume the θ_1 and θ_2 are correlated, denoted the baseline and exit latent traits for the repeated measures. The distribution function is [1],

$$P(X = x|\theta) = \frac{\exp(r_1\theta_1 + r_2\theta_2 + \sum_i \eta_{ix_i})}{K_1(\theta_1)K_2(\theta_2)} \tag{2}$$

where $r_1 = \sum_{i\in I_1} x_i$ and $r_2 = \sum_{i\in I_2} x_i$, which are sufficient statistics for θ_1 and θ_2 in Rasch model. The joint likelihood is [1],

$$l(\eta_i, \theta_1, \theta_2) = r_1\theta_1 + r_2\theta_2 + \sum_i \eta_{ix_i} - \log K_1(\theta_1) - \log K_2(\theta_2) \tag{3}$$

Under the assumption,

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$$

The marginal log-likelihood for person v is [1],

$$l(\eta_i) = \sum_i \eta_{ix_i} + \log \int_{\mathscr{R}}\int_{\mathscr{R}} \frac{\exp(r_1\theta_1 + r_2\theta_2)}{K_1(\theta_1)K_2(\theta_2)} \varphi(\theta_1, \theta_2) d\theta_1 d\theta_2 \tag{4}$$

In the algorithm, only the change $\mu = \mu_2 - \mu_1$ is estimates with the assumption, Under the assumption,

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$$

Under this assumption, only the change between the two points is estimated. The correlation between these longitudinal latent traits should be high and positive. The variance covariance matrix is estimated in the model.

3. Results

The parameters were estimated based on the restriction that there were no item parameters that changed over time. Under this setting, the items of different recall intervals evaluate the adherence the same at different time points. The results were shown in Table 2. The parameters were estimated at baseline and exit equally in the model. We use the self-reported adherence data in Multi-site Adherence Collaboration in HIV among 14 sites in United States. Eligible studies were required to have: (1) a longitudinal study design with at least 3 repeated measurements; (2) MEMS adherence data; (3) VL and clinical outcomes; and (4) psychosocial and behavioral measures. The details description of the 16 HIV adherence studies [24] from 14 different institutions or universities are described in other papers.

Table 1: My caption

	Category	1-day	2-day	3-day	4-day	7-day
Baseline	**<50%**	103 (6.4%)	97 (6.1%)	101 (6.3%)	25 (1.6%)	4 (0.2%)
	50-85%	102 (6.4%)	90 (5.6%)	86 (5.4%)	8 (0.5%)	20 (1.2%)
	85-99%	9 (0.6%)	15 (0.9%)	7 (0.4%)	(0%)	34 (2.1%)
	100%	1390 (86.8%)	1282 (80.1%)	1283 (80.1%)	389 (24.3%)	44 (2.7%)
Exit	**<50%**	136 (8.5%)	119 (7.4%)	114 (7.1%)	29 (1.8%)	14 (0.9%)
	50-85%	97 (6.1%)	84 (5.2%)	74 (4.6%)	8 (0.5%)	16 (1%)
	85-99%	9 (0.6%)	15 (0.9%)	6 (0.4%)	(0%)	28 (1.7%)
	100%	1362 (85.1%)	1270 (79.3%)	1296 (80.9%)	385 (24%)	44 (2.7%)

3.1 Adherence

The self-reported adherence at baseline week and exit week is converted into the following response categories,

- 0 if the self-reported adherence $\leq 50\%$
- 1 if the self-reported adherence $(50\%, 85\%)$
- 2 if the self-reported adherence $\geq 85\%$, but not perfect
- 3 if the self-reported adherence is perfect 100%

Patients with adherence at different levels require different level of care and intervention engagement. For those who have adherence level less than 50%, an intensive intervention might be necessary to boost up adherence and prevent these patients from treatment failure. For those who are less than 85% are more likely to develop drug resistance. They may not benefit from the ARVs and might need close attention on the type of regimens. For those who maintain adherence level over 85% might have the best clinical outcome but also more vulnerable to viral mutation. Different level of adherence may involve various level of care management. From the results of Figure 1 and Figure 2, the weekly recall of adherence has more clear cut off at the these thresholds. Therefore, if the scientific focus is to identify those who are needs intensive care management on both adherence and treatment engagement, and those who are more likely to develop drug resistance, we suggest the recall interval should be as long as seven days to distinguish the variation. The self-reported adherence data for 1601 subjects is summarized as table 1,

3.2 Threshold parameters

The parameter estimation is as shown in table 2. We preset the item parameters at baseline and exit to be equal. The only difference is between the personal parameter will change - the PRO adherence. The item characteristic curves are shown in the figure 1 and figure 2. The weekly recall can identify the different level of categories better compared with other recall intervals.

3.3 Parameter Estimation

The person level parameter estimation is shown in table 3. Overall the adherence increased when exit from the study, compare their baseline self-reported level. However, the change is not significant. The latent variables at two different time points θ_1 and θ_2 are highly correlated. The variances for baseline latent θ_1 and exit θ_2 are very similar.

Table 2: Original thresholds for all items

Baseline	**Exit**	**Estimate**	**SE**	Lower	Upper
bsSR1—1	exSR1—1	-2.08	0.2	-2.39	-1.76
bsSR1—2	exSR1—2	0.93	0.3	0.4	1.46
bsSR1—3	exSR1—3	-5.41	0.3	-5.9	-4.92
bsSR2—1	exSR2—1	-2.02	0.2	-2.33	-1.7
bsSR2—2	exSR2—2	0.31	0.2	-0.13	0.76
bsSR2—3	exSR2—3	-4.8	0.2	-5.19	-4.4
bsSR3—1	exSR3—1	-1.96	0.2	-2.28	-1.64
bsSR3—2	exSR3—2	1.04	0.3	0.44	1.65
bsSR3—3	exSR3—3	-5.67	0.3	-6.24	-5.1
bsSR4—1	exSR4—1	-0.65	0.3	-1.29	-0.01
bsSR4—2	exSR4—2	-4.29	0.3	-4.83	-3.76
bsSR7—1	exSR7—1	-3	0.4	-3.71	-2.29
bsSR7—2	exSR7—2	-1.77	0.3	-2.29	-1.25
bsSR7—3	exSR7—3	-0.23	0.2	-0.7	0.23

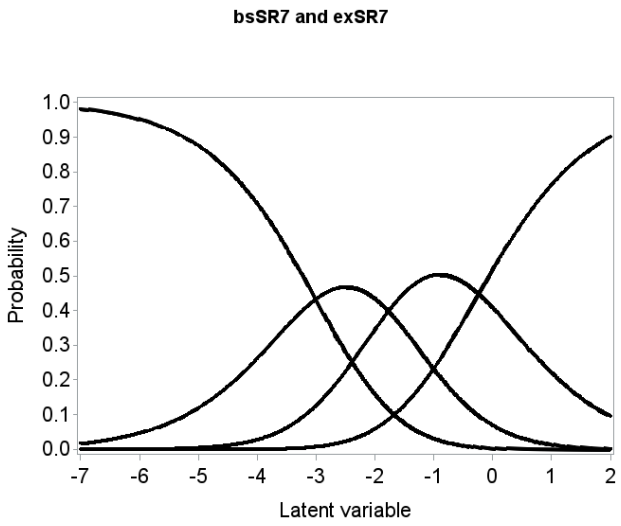
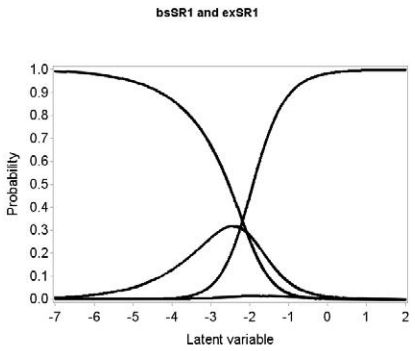


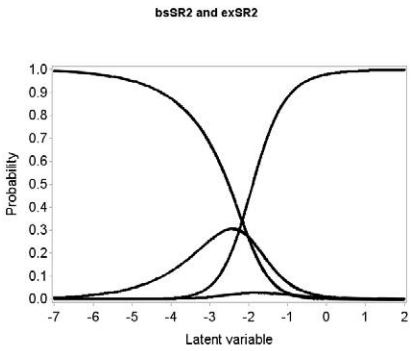
Figure 1: Seven days (weekly) recall ICC

Table 3: Change in latent mean, latent correlation and variances

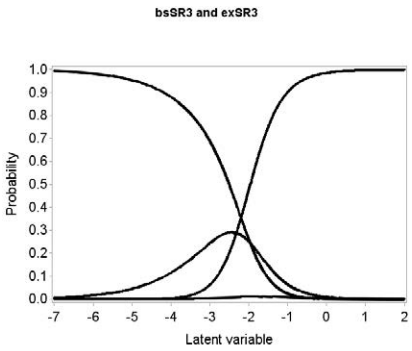
Parameter	**Estimate**	**Standard Error**	Lower	Upper
μ	0.109	0.119	-0.124	0.343
ρ	0.832	0.021	0.790	0.873
σ_1	1.548	0.089	1.374	1.722
σ_2	1.779	0.104	1.574	1.983



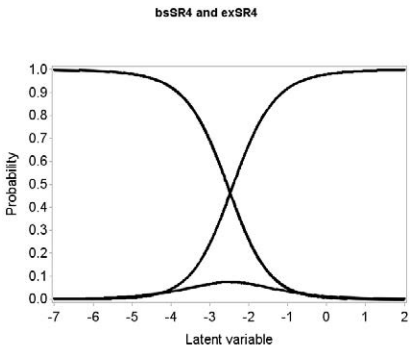
(a) 1 day recall



(b) 2 days recall



(c) 3 days recall



(d) 4 days recall

Figure 2: One to four days recall ICCs

4. Discussion

Self-reported adherence is easy to use and administrative in the clinical settings. However, it is long term been considered over-estimated and less accurate compared to adherence subjectively measured by device, such as MEMS or Wisepill. There was few recommendations for using the recall intervals based on statistical methods for thresholds.

One recommendation from this paper is to use weekly recall of self-reported adherence in the study aimed to identify those with lower adherence level that may lead to treatment failure, those who need some assistance maintain a relative high adherence, those who have relatively high adherence level but are vulnerable to viral mutation resistance, and those who are in perfect self-care.

Currently, there are different opinions for the optimal recall interval to access self-reported adherence. Commonly used intervals are one day recall, three-day recall, and weekly recall. In this paper, we recommend the weekly recall in order to have better threshold for different level of adherence. The precision and variation of self-reported adherence level over time are both important when using to access medication taking behavior.

In the longitudinal IRT models, we had heavier burden on number of unknowns. The parameters we estimated included the item parameters related to the responses of each survey items and latent trait related to individuals. These parameters could change over time or could be the same over time. In public health, the latent trait was usually the disease related measure. We mostly interested the change of this disease related θ change over time, rather than the item changes. Therefore, in most cases we assumed, the parameters should be the same over time.

We did not evaluate the details of person parameter in this paper. It should be easily implemented using regular Rasch model outputs of threshold parameters to estimate the person level traits [22, 25]. The longitudinal version with more than two time points should be the next step to evaluate the changes overtime as mentioned in the paper [1]. The constraints between of parameters shift are of challenge. There is no guideline on the set ups for these constraints so far.

In this report, we did not talk about the missing value strategy. The missing in each category, may either due to study design or not available from the patients. There are other strategies of handling systematically missingness due to study design. We did not go to that level of details here.

5. Acknowledgments

We would like to express our great appreciation to MACH14 investigators and MACH14 patients for their contribution to the rich data set.

References

[1] M. Olsbjerg and K. B. Christensen, "maximum likelihood estimation in longitudinal polytomous rasch models," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–24, 2015.

[2] W. J. van der Linden and R. K. Hambleton, *Handbook of modern item response theory*. Springer Science and Business Media, 2013.

[3] K. S. Chan, A. L. Gross, L. E. Pezzin, J. Brandt, and J. D. Kasper, "Harmonizing measures of cognitive performance across international surveys of aging using item response theory," *Journal of aging and health*, vol. 27, no. 8, pp. 1392–1414, 2015.

[4] D. Atkins, S. M. Chang, G. Gartlehner, D. I. Buckley, E. P. Whitlock, E. Berliner, and D. Matchar, "Assessing applicability when comparing medical interventions: Ahrq and the effective health care program," *Journal of clinical epidemiology*, vol. 64, no. 11, pp. 1198–1207, 2011.

[5] A. L. Gross, R. N. Jones, T. G. Fong, D. Tommet, and S. K. Inouye, "Calibration and validation of an innovative approach for estimating general cognitive performance," *Neuroepidemiology*, vol. 42, no. 3, pp. 144–153, 2014.

[6] R. N. Jones and S. J. Fonda, "Use of an irt-based latent variable model to link different forms of the ces-d from the health and retirement study," *Social Psychiatry and Psychiatric Epidemiology*, vol. 39, no. 10, pp. 828–835, 2004.

[7] D. Liu, X. Wang, F. Pan, P. Yang, Y. Xu, X. Tang, J. Hu, and K. Rao, "Harmonization of health data at national level: a pilot study in china," *International journal of medical informatics*, vol. 79, no. 6, pp. 450–458, 2010.

[8] S. Gupta and A. Goren, "Application of item response theory in validating the morisky medication adherence scale in patients with hypertension," *Value in Health*, vol. 16, no. 3, p. A4, 2013.

[9] S. Karve, M. A. Cleves, M. Helm, T. J. Hudson, D. S. West, and B. C. Martin, "Good and poor adherence: optimal cut-point for adherence measures using administrative claims data," *Current medical research and opinion*, 2009.

[10] S. C. Kalichman and T. Grebler, "Stress and poverty predictors of treatment adherence among people with low-literacy living with hiv/aids," *Psychosomatic medicine*, vol. 72, no. 8, p. 810, 2010.

[11] J.-J. Parienti, M. Das-Douglas, V. Massari, D. Guzman, S. G. Deeks, R. Verdon, and D. R. Bangsberg, "Not all missed doses are the same: sustained nnrti treatment interruptions predict hiv rebound at low-to-moderate adherence levels," *PloS one*, vol. 3, no. 7, p. e2783, 2008.

[12] BusinessDictionary.com., "harmonization.," 2016.

[13] BioSHaRE, "Manuscript on model(s) for harmonizing longitudinal data (deliverable 3.6)," 2015.

[14] A. Sheth, "Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies," in *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pp. 2–2, IEEE, 2014.

[15] L. Griffith, E. van den Heuvel, I. Fortier, S. Hofer, P. Raina, N. Sohel, H. Payette, C. Wolfson, and S. Belleville, "Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis," *Agency for Healthcare Research and Quality (US)*, 2013.

[16] S. Van Buuren, S. Eyres, A. Tennant, and M. Hopman-Rock, "Improving comparability of existing data by response conversion," *Journal of Official Statistics*, vol. 21, no. 1, pp. 53–72, 2005.

[17] N. Minicuci, M. Noale, C. Bardage, T. Blumstein, D. J. Deeg, J. Gindin, M. Jylh, S. Nikula, A. Otero, and N. L. Pedersen, "Cross-national determinants of quality of life from six longitudinal studies on aging: the clesa project," *Aging clinical and experimental research*, vol. 15, no. 3, pp. 187–202, 2003.

[18] J. Ma, P. Raina, and L. Griffith, "Comparison of statistical harmonization methods in individual participant data meta-analysis," August 5, 2013 2013.

[19] P. Anselmi, G. Vidotto, O. Bettinardi, and G. Bertolotti, "Measurement of change in health status with rasch models," *Health and quality of life outcomes*, vol. 13, no. 1, p. 1, 2015.

[20] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.

[21] D. Andrich and G. Luo, "Conditional pairwise estimation in the rasch model for ordered response categories using principal components," *Journal of applied measurement*, vol. 4, no. 3, pp. 205–221, 2002.

[22] K. B. Christensen, "Conditional maximum likelihood estimation in polytomous rasch models using sas," *ISRN Computational Mathematics*, vol. 2013, 2013.

[23] M. Olsbjerg and K. Christensen, "Local dependence in longitudinal irt models," tech. rep., Research Report 2, Department of Statistics, University of Copenhagen, 2014.

[24] H. Liu, I. B. Wilson, K. Goggin, N. Reynolds, J. M. Simoni, C. E. Golin, M. I. Rosen, R. Gross, G. Wagner, and R. H. Remien, "Mach14: a multi-site collaboration on art adherence among 14 institutions," *AIDS and Behavior*, vol. 17, no. 1, pp. 127–141, 2013.

[25] Y. Wang, H. M. Crane, and H. Liu, "Conditional maximum likelihood rasch model in data harmonization," in *Joint Statistical Meetings 2014, Boston MA*, pp. 464–473, American Statistics Association, 2014.