Assessment of Score- and Rasch-Based Methods for Group Comparison of Longitudinal Patient-Reported Outcomes with Intermittent Missing Data (Informative and Non-Informative).

Élodie de Bock, Jean-Benoit Hardouin, Myriam Blanchin, Tanguy Le Neel, Gildas Kubis, Véronique Sébille

Assessment of score and Rasch-based methods for group comparison of longitudinal Patient-Reported Outcomes with intermittent missing data (informative and non-informative)

Author's names: Élodie de Bock[1], Jean-Benoit Hardouin[1], Myriam Blanchin[1], Tanguy Le Neel[1], Gildas Kubis[1], Véronique Sébille[1].
Affiliation ([1]): EA4275 - SPHERE
'Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences'
Faculté de Pharmacie – Université de Nantes
Pres Université Nantes Angers Le Mans
1, rue Gaston Veil – 44035 NANTES Cedex 01 France

Corresponding author: Élodie de Bock
Email: elodie.debock@univ-nantes.fr
Phone/**Fax**: +33 (0)2 40 41 29 96

Abstract

Purpose

The purpose of this study was to identify the most adequate strategy for group comparison of longitudinal PROs in the presence of possibly informative intermittent missing data. Models coming from CTT and IRT were compared.

Methods

Two groups of patients' responses to dichotomous items with three times of assessment were simulated. Different cases were considered: presence or absence of a group effect and/or a time effect, a total of 100 or 200 patients, 4 or 7 items, and two different values for the correlation coefficient of the latent trait between two consecutive times (0.4 or 0.9). Cases including informative and non-informative intermittent missing data were compared at different rates (15%, 30%). These simulated data were analysed with CTT using Score and Mixed model (SM) and with IRT using Longitudinal Rasch Mixed model (LRM). The type I error, the power and the bias of the group effect estimations were compared between the two methods.

Results

This study showed that LRM performs better than SM. When the rate of missing data rose to 30%, estimations were biased with SM mainly for informative missing data. Otherwise, LRM and SM methods were comparable concerning biases. However, regardless the rate of intermittent missing data, power of LRM was higher compared to power of SM.

Conclusions

In conclusion, LRM should be favored when the rate of missing data is higher than 15%. For other cases, SM and LRM provide similar results.

Keywords: IRT, CTT, Rasch models, longitudinal data, PROs, missing data

Introduction

Nowadays, in clinical studies, it is frequent to assess health related quality of life (QoL) using Patient Reported Outcomes (PROs) [1]. PROs allow evaluating patients' perceptions using patients' responses to items grouped into one or more dimensions of a questionnaire. Thus, PROs allow measuring subjective concepts like QoL that cannot be directly observed and such data are often called latent variables for this reason [2].

Data are usually collected in two or more groups of patients in order to compare the impact of treatments on QoL, for instance. For example, a randomized trial could be set up in order to test whether a new treatment improves patients' quality of life compared to a standard therapy. In this case, the group effect would assess the difference of quality of life levels between the two groups. The impact of other variables on quality of life such as gender, age could also be considered. Moreover, patients are frequently followed over time in order to assess the evolution of quality of life, for instance. In that case, collected data are longitudinal. Missing data are frequently encountered in longitudinal studies and can seriously impact the results with a potential loss of power and biased estimates [3, 4]. The most reliable methodological approach to handle longitudinal PROs with missing data is still under debate.
Indeed, according to the type of missing data (informative or non-informative), consequences on conclusions may be different. Mechanisms of missing data were defined by Little and Rubin [5]. They described non-informative missing data, which combine MCAR (Missing Completely At Random) and MAR (Missing At Random) data, and informative missing data called MNAR (Missing Not At Random) data. Missing data are non-informative when the probability to have a missing value either depends on the observed data (MAR) or is independent of all previous, current and future assessments (MCAR). An example of MCAR data could be forgetting to answer to an item. MAR data may correspond to the case where a response to an item is only required when a positive response has already been given to a previous item. For instance, if the response is "yes" for a given question, the patient goes directly to the next items; if the answer is "no", no other responses are required [6]. Unlike the two latter cases, if the latent variable (QoL of a patient) has an impact on the occurrence of missing data, it will be informative. For instance, it has often been observed that as the QoL of a patient gets worse, his/her propensity of non-response gets higher [7].
The pattern of missing data can also vary: a whole form (questionnaire) could be missing for one patient at different times (it is called 'intermittent missing forms' [8]), or a patient could stop the study and his/her forms would then not be available after a certain point in time (it corresponds to a 'complete dropout' [9, 10]). It is also very common that a patient doesn't answer to one or more items of a questionnaire at each time [11]. This pattern is called intermittent missing items and it will be studied in the present paper.

The Classical Test Theory (CTT) and the Item Response Theory (IRT) are the two main analytic approaches that are usually performed for PROs data analysis. CTT is based on the observed scores (possibly weighted sum of patients items' responses, interpreted as being close to the true score), while IRT, and more particularly Rasch models, links the items responses to a latent parameter (i.e. the latent trait, interpreted as the true individual QoL, for example) by a response model.
In the framework of longitudinal data and for complete data case, both approaches obtained similar results (unbiased and good power) [12]. In the same framework and for complete dropout case, both methods engendered poor power and biased estimates in case of MNAR data [13]. Moreover, for longitudinal data, in presence of possibly informative intermittent missing data, IRT appeared to be more powerful than CTT for identifying and quantifying a time effect in a single group of patients [14]. A simulation study of group comparison in a cross-sectional framework without missing data has shown that IRT performed better than CTT [15] regarding power.
The relative performance of CTT and IRT for identifying and estimating a group effect in the framework of longitudinal PROs data with possibly informative intermittent missing items is unknown and remains to be identified. The aim of the present study was to compare two methods based on CTT and IRT approaches in the context of clinical studies where longitudinal data are gathered to compare two groups of patients. Moreover, the goal is also to find the most reliable methodological approach to handle longitudinal PROs with intermittent missing items. In order to assess and compare the performance of score (CTT) and Rasch-based (IRT) methods, a simulation study was developed.

Method

A simulation study was favored to assess and compare the performance of score-based (CTT approach) and of Rasch-based (IRT approach) models. It consists in simulating datasets using a priori chosen parameters' values corresponding to different types of situations encountered in clinical studies (several sample sizes, number of

items of the questionnaire…) and in analyzing these simulated datasets using models for which appropriateness and accuracy can be measured and compared. Indeed, when datasets are simulated, true values of the different parameters are known and controlled. In that case, it is possible to compare the true values (values of the simulated parameters) and the estimated parameters (estimated values of parameters obtained when analyzing the simulated datasets with the models). Conversely, real data can't be used for that purpose because true values of the different parameters aren't known in that case. Thus, simulation study is very useful to objectively compare different methods [16].

PROs simulation

The parameters that were chosen to simulate datasets represent common situations encountered in clinical research. For each case, 500 datasets were simulated. Questionnaires were assumed to have been previously validated with CTT and IRT [17] which is currently performed when it is envisaged to analyze PROs using either score- or Rasch-based approaches [18-20]. Questionnaires contained 4 or 7 dichotomous items. The items difficulties were regularly distributed and defined by the vectors (-1; -0.5; 0.5; 1) and (-1.5; -1; -0.5; 0; 0.5; 1; 1.5) for respectively the 4-item questionnaire and the 7-item questionnaire. Two groups with 50 patients each (total sample size of 100 patients) or with 100 patients each (total sample size of 200 patients) were simulated and three times of assessment were assumed. The latent trait vector (Θ) followed a multivariate normal distribution. The mean of this distribution was $\mu_\theta = (\mu_\theta^{(1)}, \mu_\theta^{(2)}, \mu_\theta^{(3)})'$ and the covariance matrix had a first-order autoregressive structure. Concerning the correlation coefficient of the latent trait between two consecutive times, two different values were used: 0.4 or 0.9. Monte Carlo simulations with a longitudinal Rasch model were used to simulate the patients' responses. Two assumptions regarding group effect were simulated: presence of a group effect or no. When a group effect was assumed, it was equal to 0.5 for the latent trait and corresponded to 0.38 and 0.63 for the score (difference between the means of the scores) for the 4-item questionnaire and the 7-item questionnaire, respectively; or 0.2 for the latent trait and corresponded to 0.15 and 0.25 for the score for the 4-item questionnaire and the 7-item questionnaire, respectively [13]. A group effect equals to 0.5 means that a difference between the means of the latent trait of each group is equal to 0.5. It means that one of the groups is composed of patients who have a better QoL (on average 0.5 points) than patients of the other group. In other words, when a group effect exists, at each time of assessment: $\mu_{\theta,group2}^{(t)} - \mu_{\theta,group1}^{(t)} = 0.5$ ($\mu_{\theta,group1}^{(t)}$ and $\mu_{\theta,group2}^{(t)}$ are the means of the latent trait at time t for the patients of the group 1 and 2, respectively). Moreover, a time effect between two consecutive measures was simulated as equal to 0.2; or, equal to 0, when no time effect was assumed. It means that, globally, for all patients of the two groups, there is a difference of 0.2 for the latent trait (QoL) between two consecutive measures that is to say that QoL increases with time. In other words, when a time effect exists: $\mu_\theta^{(2)} - \mu_\theta^{(1)} = \mu_\theta^{(3)} - \mu_\theta^{(2)} = 0.2$ ($\mu_\theta^{(1)}$, $\mu_\theta^{(2)}$ and $\mu_\theta^{(3)}$ are the means of the latent trait for the patients of the two groups at the first, second, and third time of assessment, respectively).

The intermittent missing items were created from the complete simulated datasets using a logistic model [13]. A second latent variable was defined and represented the propensity of non-response that can vary according to patients. MCAR items were simulated with a correlation coefficient between the latent variable of interest (QoL for instance) and the patient's propensity of non-response equal to 0. Indeed, non-informative missing items are assumed to be independent of the latent trait. For MNAR items, informative missing data are assumed to depend on the latent trait. So, in that case, this correlation coefficient was different from 0. Moreover, we assumed that, patients tend to answer to items less often when their QoL is deteriorated. In that case, the value of this coefficient is negative. Therefore, for MNAR items, this correlation coefficient was equal to -0.9. Two rates of intermittent missing items were simulated: π = 15% or 30%. It means that, for all patients and times combined, there was 15% or 30% on average of intermittent missing data for each item. Hence, the rate of intermittent missing items may vary according to the patient, the time of assessment, the patients' group, and the level of QoL considered.

PROs analysis

Due to the longitudinal design of the simulated study, datasets were analyzed with a score-based model that will be named Score and Mixed model (SM) for the CTT approach and with a Rasch-based model that will be called longitudinal Rasch Mixed model (LRM) for the IRT approach. When data are simulated using a Rasch model, the assumptions of CTT and IRT approaches are verified [17]. Hence, both approaches can be applied to the simulated datasets.

<u>SM method</u>

The CTT approach is based on the computation of a score. With this approach, the "true" score, which is a representation of the studied concept (true QoL of the patient, for example), is assumed to be well estimated by the observed score (observed QoL of the patient) [2].
The SM method analyzes the patients' scores at each time. The observed score for a patient at a given time is obtained by summing the responses of this patient to all items at this time. Then, a linear mixed model is applied on the observed scores, calculated at each time of assessment, in order to test whether a group effect and a time effect exist. The null hypothesis for the group effect is: $\mu_{score,group1} = \mu_{score,group2}$ ($\mu_{score,group1}$ and $\mu_{score,group2}$ are the means of the scores for the patients of the group 1 and 2, respectively). The null hypothesis for the time effect is: $\mu_{score}^{(1)} = \mu_{score}^{(2)} = \mu_{score}^{(3)}$ ($\mu_{score}^{(1)}$, $\mu_{score}^{(2)}$ and $\mu_{score}^{(3)}$ are the means of the scores for the patients of the two groups at the first, second, and third time of assessment, respectively).
All items must be answered by the patient in order to compute a score. Indeed, if at least one item is missing, the computation of the score cannot be made, which is an issue with incomplete data. It is recommended by several scoring manuals of questionnaires (SF-36, QLQ-C30, etc.) to impute a missing data by the mean response of the patient to the other items. It is usually performed when the amount of missing items at a given time does not exceed 50% for a given patient [21]. In other cases, the score cannot be computed. Thus, imputation allows decreasing the rate of missing values by increasing the number of computed scores. This method is known as Personal Mean Score (PMS) imputation [22] and it was applied before using the SM method. For the simulated datasets, data of patients with one and three missing items maximum were imputed for the 4-item questionnaire and the 7-item questionnaire, respectively in order to respect the fixed limit of 50% of imputed items.

The Proc MIXED of SAS was used with the Restricted Maximum Likelihood (REML) estimation in order to estimate and assess the significance of the parameters of the model [23].

LRM method

IRT explains the probability of a response to an item as a function of the latent trait level (true QoL for a patient, for instance) and items' parameters (difficulties). The item difficulty is an item characteristic. The lower the item difficulty, the higher the probability of positive answer (favorable response of the patient to this item regarding the latent trait being measured).
The LRM model is a longitudinal Rasch model [24-26, 12]: the items' responses and the latent variable are connected by a logistic link function.
The specific objectivity property of the Rasch model allows obtaining consistent estimations of the parameters associated to the latent trait independently from the items used for these estimations. Indeed, even if some patients do not respond to all items, estimates of the latent trait's parameters should be unbiased.

The null hypothesis for the group effect is: $\mu_{\theta,group1} = \mu_{\theta,group2}$ ($\mu_{\theta,group1}$ and $\mu_{\theta,group2}$ are the means of the latent trait for the patients of the group 1 and 2, respectively). The null hypothesis for the time effect is: $\mu_{\theta}^{(1)} = \mu_{\theta}^{(2)} = \mu_{\theta}^{(3)}$ ($\mu_{\theta}^{(1)}$, $\mu_{\theta}^{(2)}$ and $\mu_{\theta}^{(3)}$ are the means of the latent trait for the patients of the two groups at the first, second and third time of assessment, respectively).

The Gllamm module of Stata was used to estimate and assess the significance of the parameters of the model [27].

For SM and LRM analyses, an unstructured covariance matrix was used assuming that all covariances and variances parameters can be different between times of assessments.
Figure 1 illustrates LRM and SM implementations using Stata and SAS, respectively.

Criteria for methods' comparison

The bias of the group effect estimations, the type I error and the power of the tests were evaluated in order to compare both methods.

- **Concerning the potential bias of group effect estimations, for convenience, the means of the group effect estimations (means obtained on the 500 datasets of each case using SM or LRM analyses) were compared to the theoretical true value (simulated value) using a t-test.**
- **The type I error corresponded to the proportion of rejection of the null hypothesis of group effect on the 500 datasets of each case where no group effect had been simulated. For the type I error, the expected rate was 5%.**

- **The power corresponded to the proportion of rejection of the null hypothesis of group effect on the 500 datasets of each case where a group effect had been simulated. We considered that a power difference of 0.05 between two methods could be considered as relevant.**

Results

The tables 1 to 3 give the results obtained on simulated datasets regarding bias when no group effect was simulated and bias when a group effect was simulated, type I error and power**,** respectively.

Complete datasets (results for π=0% in table 1 and results "complete data" in tables 2 and 3)

Regarding type I error and power, SM and LRM methods displayed similar results whatever the values of the parameters. The type I errors were close to the expected value (5%). Group effect estimations were unbiased for both methods.

Intermittent missing items (results "MCAR" and "MNAR" in all tables)

Table 1 (part "No group effect") shows results of the group effect estimation when no group effect was simulated. The rate of biased values observed for LRM and SM methods was weak. Indeed, only 3 biased values over 64 values (4.7%) were noticed for LRM method and 2 over 64 (3.1%) for SM method.
Table 1 (part "Group effect") shows results of the group effect estimation when a group effect was simulated. There were 10 biased values over 64 values (16%) for LRM method and 20 over 64 (31%) for SM method. Biases appeared more often for datasets including missing data created by an MNAR mechanism. Indeed, for MNAR data, 19 biased values over the 30 biased values (63%) occurred for both LRM and SM methods. More specifically, 13 and 6 biased values over the 20 and 10 biased values (65% and 60%) occurred for SM and LRM, respectively. Moreover, estimations were more often biased with a rate of 30% of intermittent missing items compared to those with a rate of 15% of missing data. Indeed, 19 biased values over the 30 biased values (63%) concerned data with 30% of missing data for both LRM and SM methods. More specifically, 14 and 5 biased values over the 20 and 10 biased values (70% and 50%) occurred for SM and LRM, respectively.

Table 2 shows results for the type I error. The minimum value for the type I error was 3.2%; its average was at 5% and its maximum at 9.4%. The expected value of 5% was not included in the 95% confidence interval for only one type I error over 160 (0.6%). None of the parameters used for datasets simulation (number of patients and items, value of time effect, correlation of the latent trait between two consecutive times, and the type of missing items and its rate) seemed to have an impact on the type I error. Hence, the type I error was controlled for both methods.

Table 3 shows results for the power of group effect tests. Some powers must be interpreted with caution because the associated group effect estimations were biased. For both methods, the number of patients and items as well as the correlation between two consecutive times impacted power. As expected, power decreased with the reduction of sample size, and it increased with the number of items. It could be noticed that the observed power was lower when the correlation of the latent trait between two consecutive times was higher. No variation could really be explained by the type of intermittent missing items (MCAR or MNAR). However, power decreased when the rate of intermittent missing items increased. Whatever the parameters and the type of intermittent missing items, power was usually greater for LRM method compared to SM method. We could observe that 17 cases (corresponding to datasets with 30% of intermittent missing items) over 64 (27%) had a power difference equal or superior to 0.05 for LRM compared to SM. Conversely, a power difference equal or superior to 0.05 for SM compared to LRM was never observed whatever the considered case. Graph 1 illustrates this fact: whatever the type of intermittent missing items, power was higher for LRM as compared to SM method. Moreover the difference between the power obtained with the LRM method as compared to the SM method was greater in case of 30% of intermittent missing items as compared to 15%.

Concerning datasets simulated with a group effect equal to 0.2 (results not shown**),** results were similar except for the group effect estimations when a group effect was simulated. Indeed, in that case, estimations were unbiased for both methods.

Discussion

In clinical studies, QoL has become a criterion of interest often assessed using PROs. The evolution of this criterion is often assessed over time in different groups of patients. In this kind of framework, intermittent missing items are commonly encountered. If missing items are linked to the patient's QoL (MNAR data), they can seriously impact the conclusions and may be truly problematic. The purpose of the present study was to compare the appropriateness and accuracy of two models (SM and LRM) based on CTT and IRT approaches to detect and quantify a group effect on longitudinal PROs data with possibly informative intermittent missing items.

Concerning complete cases, results were very close to those obtained by Blanchin et al. [12] with controlled type I error and similar power for SM and LRM. For cases with intermittent missing items, LRM appeared to be more powerful than SM. Moreover, for 27% of the simulated cases a power difference between both methods equal or superior to 0.05 in favor of LRM was observed. Concerning biases, when the rate of missing data rose to 30%, estimations were biased mainly with SM for informative missing data. Otherwise, LRM and SM methods were comparable concerning biases.
It could be highlighted that, with a group effect equal to 0.2, conclusions were very similar to those obtained with a group effect equal to 0.5 except for estimations' biases. Indeed, no bias was noticed. It could be explained by the way that data were simulated. For MNAR data, the worse the QoL, the higher the rate of missing data. Hence, with a difference of QoL equal to 0.5 between groups, the rate of missing data is more differentiated between groups than with a group effect equal to 0.2. The unbalance of missing data rates between groups engendered biases on group effect estimations.
This study also brought out a known impact of the type of missing data on estimations that were more often biased for informative missing items (MNAR data) than for non-informative missing items (MCAR data).

Concerning SM, a PMS imputation was applied because this imputation is commonly used for well-known questionnaires (SF-36, QLQ-C30, etc.) for practical reasons (easy to use and to obtain a score for one patient with only data of this patient) [28]. Other methods like multiple imputations could improve results obtained with SM. It would be interesting to look for the impact that could have these methods on this type of data. For LRM, power was overall higher than the one obtained with SM and no imputation was necessary. Moreover, LRM appeared to be less often biased compared to SM. The specific objectivity property of the Rasch model could explain the difference between the results observed for Rasch-based model (LRM) and score-based model (SM).

In this study, the simulated group effect was constant over time and it could be a limitation. Indeed, most clinical studies are randomized. Consequently, at the first time of assessment, no group effect is supposed to exist and a group effect can appear afterwards and may be different depending on the time of assessment. It could be interesting to complete this study with other simulation cases including an interaction between the time effect and the group effect. The hypothesis could be made that results might be close to those obtained in this study because no assumption concerning the linearity of effects was necessary to analyze data with both methods.

Concerning the MNAR data, we considered that a patient tends to be too tired to respond when her/his QoL is deteriorated. However, it is possible to do the reverse hypothesis. Indeed, a patient with a good QoL could decide that it is not necessary to answer to all items because she/he feels well. In this case, the correlation between the latent variable of interest (QoL for instance) and the patient's propensity of non-response would be positive. Hence, the decrease of the QoL level would reduce the rate of missing data. In this study, we studied the case where the increase of the QoL level reduced the rate of missing data. We could do the hypothesis that the methods SM and LRM would perform similarly for both cases.

These results showed that LRM performed better than SM to assess a group effect in the framework of a longitudinal study with possibly informative intermittent missing items. However, it should be mentioned that when the rate of intermittent missing data was lower or equal to 15%, SM could provide unbiased estimations and display a power that was close to LRM. **In conclusion, since LRM is more difficult to implement than SM, the small increase in performance when there are few missing data (e.g. 15%, according to the simulation plan) does not warrant the increased effort in using LRM. However, when there is considerable missing data (e.g. 30%, regardless of the type of missing, according to the simulation plan), LRM is to be seriously considered as SM is biased.** In that case, when the implementation of LRM is necessary, it is possible to do it with Stata or SAS using the Gllamm module of Stata (as indicated on Figure 1) or the PROC NLMIXED of SAS [14].

Acknowledgments

This study was supported by the Ligue Nationale Contre le Cancer and the Comité de Loire-Atlantique de la Ligue Contre le Cancer.

References

[1] Garcia, S.F., Cella, D., Clauser, S.B., Flynn, K.E., Lad, T., Lai, J.S., Reeve, B.B., Smith, A.W., Stone, A.A., Weinfurt, K. (2007). Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. doi:10.1200/JCO.2007.12.2341.

[2] Falissard B. (2001). *Mesurer la subjectivité en santé : Perspective méthodologique et statistique*. Paris: Masson.

[3] Fairclough, D.L., Peterson, H.F., Chang, V. (1998). Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine*. doi:10.1002/ (SICI)1097-0258(19980315/15)17:5/7<667::AID-SIM813>3.0.CO;2-6.

[4] Bernhard, J., Cella, D.F., Coates, A.S., Fallowfield, L., Ganz, P.A., Moinpour, C.M., Mosconi, P., Osoba, D., Simes, J., Hürny, C. (1998) Missing quality of life data in cancer clinical trials: serious problems and challenges. *Statistics in Medicine*. doi:10.1002/(SICI) 1097-0258(19980315/15)17:5/7<517::AID-SIM799>3.0.CO;2-S.

[5] Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience. Second Edition.

[6] Schafer, J.L., Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*. doi:10.1037/1082-989X.7.2.147.

[7] Holman, R., Glas, C.A.W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/j.2044-8317.2005.tb00312.x.

[8] Curran, D., Molenberghs, G., Fayers, P.M., Machin, D. (1998). Incomplete quality of life data in randomized trials: missing forms. *Statistics in Medicine*. doi:10.1002/(SICI)1097-0258(19980315/15)17:5/7<697::AID-SIM815>3.0.CO;2-Y.

[9] Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431), 1112-1121.

[10] Curran, D., Bacchi, M., Schmitz, S.F.H., Molenberghs, G., Sylvester, R.J. (1998). Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine*. doi:10.1002/(SICI)1097-0258(19980315/15)17:5/7<739::AID-SIM818>3.0.CO; 2-M.

[11] Fayers, P.M., Curran, D., Machin, D. (1998). Incomplete quality of life data in randomized trials: missing items. *Statistics in Medicine*. doi:10.1002/(SICI) 1097-0258(19980315/15)17:5/7<679::AID-SIM814>3.0.CO;2-X.

[12] Blanchin, M., Hardouin, J-B, Le Neel, T., Kubis, G., Blanchard, C., Mirallié, E., Sébille, V. (2011). Comparison of CTT and rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Statistics in Medicine*. doi:10.1002/sim.4153.

[13] Blanchin, M., Hardouin, J-B., Le Neel, T., Kubis, G., Sébille, V. (2011). Analysis of longitudinal patient reported outcomes with informative and non-informative dropout: Comparison of CTT and rasch-based methods. *International Journal of Applied Mathematics and Statistics*, 24(SI-11A), 107-124.

[14] de Bock, E., Hardouin, J-B., Blanchin, M., Le Neel, T., Kubis, G., Dantan, E., Bonnaud-Antignac, A., Sébille, V. (2013). Rasch-family models are more valuable than score-based approaches for analysing longitudinal PRO with intermittent missing data. *Statistical Methods in Medical Research*: Published online.

[15] Hamel, J-F., Hardouin, J-B., Le Neel, T., Kubis, G., Roquelaure, Y., Sébille, V. (2012). Biases and Power for Groups Comparison on Subjective Health Measurements. *PLoS ONE*. doi:10.1371/journal.pone.0044695.

[16] Burton, A., Altman, D.G., Royston, P., Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*. doi:10.1002/sim.2673.

[17] Holland, P.W., Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68(1), 123-149.

[18] Kissane, D.W., Patel, S.G., Baser, R.E., Bell, R., Farberov, M., Ostroff, J.S., Li, Y., Singh, B., Kraus, D.H., Shah, J.P. (2012). Preliminary evaluation of the reliability and validity of the shame and stigma scale in head and neck cancer. *Head & Neck*. doi:10.1002/hed.22943.

[19] Cella, D., Beaumont, J., Webster, K., Lai, J.S., Elting, L. (2006). Measuring the concerns of cancer patients with low platelet counts: the functional assessment of cancer Therapy-Thrombocytopenia (FACT-Th) questionnaire. *Supportive Care in Cancer*. doi:10.1007/s00520-006-0102-1.

[20] Bjorner, J., Petersen, M., Groenvold, M., Aaronson, N., Ahlner-elmqvist, M., Arraras, J., Brédart, A., Fayers, P., Jordhoy, M., Sprangers, M., et al. (2004). Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. *Quality of Life Research*. doi:10.1007/s11136-004-7866-x.

[21] Fayers, P.M., Machin, D. (2007). *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. John Wiley & Sons, Ltd. Second edition.

[22] Peyre, H., Leplège, A., Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the french 2003 decennial health survey. *Quality of life research*. doi:10.1007/s11136-010-9740-3.

[23] Tenenhaus, M. (1999). Statistique et logiciels : Analyse de la variance à effets mixtes utilisation de la proc MIXED : Mais que reste-t-il à la proc GLM ? *La Revue de Modulad*, (23), 53-67.

[24] Fischer, G.H., Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.

[25] Glas, C.A., Geerlings, H., van de Laar, M.A., Taal, E. (2009). Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials*. doi:10.1016/j.cct.2008.12.003.

[26] Davier, M., Meiser, T. (2007). *Rasch models for longitudinal data*. New York: Springer, Statistics for Social and Behavioral Sciences.

[27] Zheng, X., Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal*, 7(3), 313-333.

[28] Dempster, A.P., Rubin, D.B. (1983). *Overview, in Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*. New-York: Academic Press.

Table 1: Group effect estimations ($\widehat{\beta group}$) and standard deviations (s.d.) when no group effect was simulated ($\beta_{groupLRM}=0$) and when a group effect was simulated ($\beta_{groupLRM}=0.5$) for Score Mixed model (SM) with PMS imputation and Longitudinal Rasch Mixed model (LRM) methods for different values of groups' size (N), time effect, number of items (J), latent variable correlation (ρ_θ), proportion of missing data (π) and for three cases (complete, MCAR and MNAR data). Analyses performed with an unstructured covariance matrix in SM and LRM methods.

					No group effect – SM§ : 0 / LRM : 0								**Group effect – SM§ : 0.38 for J=4 ; 0.63 for J=7 / LRM : 0.5**							
					MCAR				**MNAR**				**MCAR**				**MNAR**			
					SM		**LRM**		**SM**		**LRM**		**SM**		**LRM**		**SM**		**LRM**	
N	**Time effect**	**J**	**ρ_θ**	**π**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**	$\widehat{\beta group}$	**s.d.**
100	0	4	0.4	0%	*0.009*	*0.107*	*0.012*	*0.142*	*0.009*	*0.107*	*0.012*	*0.142*	*0.377*	*0.109*	*0.501*	*0.147*	*0.377*	*0.109*	*0.501*	*0.147*
				15%	0.010	0.116	0.014	0.147	0.011	0.115	0.015	0.149	0.374	0.120	0.500	0.153	0.375	0.121	0.499	0.156
				30%	0.010	0.136	0.015	0.154	0.007	0.133	0.011	0.154	0.381	0.137	0.507	0.163	0.352	0.135	0.493	0.163
			0.9	0%	*-0.002*	*0.118*	*-0.004*	*0.158*	*-0.002*	*0.118*	*-0.004*	*0.158*	*0.382*	*0.127*	*0.511*	*0.173*	*0.382*	*0.127*	*0.511*	*0.173*
				15%	0.001	0.127	-0.001	0.164	-0.001	0.129	-0.002	0.168	0.386	0.135	0.514	0.177	0.381	0.135	0.515	0.179
				30%	0.003	0.142	-0.001	0.166	-0.001	0.134	-0.001	0.174	0.383	0.148	0.515	0.183	0.372	0.149	0.509	0.190
		7	0.4	0%	*0.004*	*0.154*	*0.003*	*0.122*	*0.004*	*0.154*	*0.003*	*0.122*	*0.621*	*0.166*	*0.494*	*0.133*	*0.621*	*0.166*	*0.494*	*0.133*
				15%	0.000	0.166	0.002	0.127	0.007	0.160	0.005	0.125	0.620	0.174	0.494	0.138	0.616	0.173	0.492	0.137
				30%	0.002	0.184	0.000	0.135	0.001	0.188	0.001	0.134	0.625	0.186	0.499	0.144	0.594	0.184	0.490	0.140
			0.9	0%	*-0.002*	*0.193*	*-0.002*	*0.156*	*-0.002*	*0.193*	*-0.002*	*0.156*	*0.623*	*0.200*	*0.501*	*0.159*	*0.623*	*0.200*	*0.501*	*0.159*
				15%	-0.002	0.199	0.000	0.157	0.003	0.199	0.000	0.160	0.624	0.199	0.503	0.162	0.622	0.208	0.501	0.166
				30%	-0.002	0.214	-0.001	0.166	-0.002	0.217	-0.004	0.167	0.628	0.217	0.498	0.168	0.607	0.221	0.497	0.172
	0.2	4	0.4	0%	*0.007*	*0.105*	*0.009*	*0.141*	*0.007*	*0.105*	*0.009*	*0.141*	*0.380*	*0.107*	*0.507*	*0.144*	*0.380*	*0.107*	*0.507*	*0.144*
				15%	0.008	0.114	0.011	0.148	0.006	0.116	0.009	0.150	0.383	0.119	0.510	0.153	0.374	0.119	0.505	0.152
				30%	0.009	0.133	0.007	0.158	0.017	0.135	0.009	0.159	0.374	0.137	0.500	0.165	0.362	0.131	0.502	0.162
			0.9	0%	*-0.003*	*0.121*	*-0.005*	*0.164*	*-0.003*	*0.121*	*-0.005*	*0.164*	*0.365*	*0.127*	*0.493*	*0.173*	*0.365*	*0.127*	*0.493*	*0.173*
				15%	-0.004	0.125	-0.006	0.169	-0.002	0.130	-0.007	0.173	0.365	0.133	0.494	0.175	0.363	0.135	0.490	0.174
				30%	0.001	0.142	0.004	0.177	-0.008	0.138	-0.007	0.174	0.361	0.142	0.491	0.179	0.352	0.143	0.493	0.181
		7	0.4	0%	*-0.003*	*0.162*	*-0.002*	*0.130*	*-0.003*	*0.162*	*-0.002*	*0.130*	*0.618*	*0.154*	*0.493*	*0.123*	*0.618*	*0.154*	*0.493*	*0.123*
				15%	-0.004	0.170	-0.004	0.133	-0.003	0.171	-0.001	0.134	0.616	0.163	0.491	0.129	0.615	0.164	0.493	0.130
				30%	-0.004	0.187	-0.001	0.142	-0.011	0.190	-0.007	0.141	0.615	0.183	0.494	0.137	0.593	0.175	0.489	0.134
			0.9	0%	*-0.001*	*0.190*	*-0.001*	*0.156*	*-0.001*	*0.190*	*-0.001*	*0.156*	*0.608*	*0.201*	*0.486*	*0.164*	*0.608*	*0.201*	*0.486*	*0.164*
				15%	-0.004	0.198	-0.002	0.160	-0.004	0.197	-0.003	0.157	0.609	0.208	0.489	0.169	0.605	0.208	0.484	0.168
				30%	-0.009	0.203	-0.004	0.162	-0.005	0.211	-0.003	0.163	0.604	0.214	0.486	0.171	0.578	0.224	0.480	0.175
50	0	4	0.4	0%	*0.001*	*0.148*	*0.002*	*0.197*	*0.001*	*0.148*	*0.002*	*0.197*	*0.376*	*0.154*	*0.501*	*0.209*	*0.376*	*0.154*	*0.501*	*0.209*
				15%	-0.002	0.161	0.001	0.209	-0.001	0.160	0.003	0.203	0.374	0.169	0.497	0.221	0.375	0.169	0.501	0.221
				30%	-0.001	0.194	0.003	0.228	-0.008	0.187	-0.010	0.223	0.370	0.203	0.499	0.234	0.362	0.191	0.499	0.230
			0.9	0%	*0.008*	*0.176*	*0.010*	*0.237*	*0.008*	*0.176*	*0.010*	*0.237*	*0.374*	*0.167*	*0.503*	*0.225*	*0.374*	*0.167*	*0.503*	*0.225*
				15%	0.006	0.187	0.008	0.245	0.008	0.188	0.016	0.248	0.371	0.177	0.505	0.233	0.371	0.179	0.506	0.236
				30%	0.012	0.209	0.009	0.257	0.003	0.213	0.015	0.263	0.361	0.195	0.496	0.243	0.362	0.207	0.497	0.248
		7	0.4	0%	*-0.014*	*0.240*	*-0.012*	*0.192*	*-0.014*	*0.240*	*-0.012*	*0.192*	*0.620*	*0.246*	*0.494*	*0.199*	*0.620*	*0.246*	*0.494*	*0.199*
				15%	-0.016	0.248	-0.010	0.196	-0.009	0.251	-0.008	0.196	0.622	0.259	0.495	0.206	0.622	0.256	0.496	0.203
				30%	-0.015	0.266	-0.012	0.201	-0.012	0.270	-0.011	0.205	0.628	0.269	0.502	0.206	0.594	0.276	0.487	0.212
			0.9	0%	*-0.014*	*0.277*	*-0.010*	*0.228*	*-0.014*	*0.277*	*-0.010*	*0.228*	*0.663*	*0.280*	*0.528*	*0.233*	*0.663*	*0.280*	*0.528*	*0.233*
				15%	-0.009	0.289	-0.007	0.233	-0.013	0.291	-0.009	0.231	0.663	0.295	0.529	0.240	0.666	0.286	0.535	0.231
				30%	-0.013	0.303	-0.011	0.234	-0.007	0.308	-0.009	0.236	0.663	0.303	0.529	0.240	0.635	0.316	0.529	0.241
	0.2	4	0.4	0%	*-0.001*	*0.149*	*-0.002*	*0.200*	*-0.001*	*0.149*	*-0.002*	*0.200*	*0.393*	*0.152*	*0.523*	*0.205*	*0.393*	*0.152*	*0.523*	*0.205*
				15%	0.002	0.162	0.000	0.210	0.001	0.164	0.000	0.210	0.393	0.172	0.523	0.215	0.388	0.163	0.519	0.212
				30%	-0.001	0.192	-0.008	0.219	0.003	0.188	0.000	0.222	0.393	0.199	0.528	0.234	0.366	0.188	0.521	0.228
			0.9	0%	*-0.005*	*0.179*	*-0.005*	*0.243*	*-0.005*	*0.179*	*-0.005*	*0.243*	*0.376*	*0.174*	*0.505*	*0.236*	*0.376*	*0.174*	*0.505*	*0.236*
				15%	-0.006	0.193	-0.007	0.257	-0.007	0.187	-0.008	0.247	0.378	0.182	0.508	0.245	0.376	0.186	0.507	0.246
				30%	-0.004	0.219	-0.012	0.266	-0.011	0.213	-0.012	0.269	0.365	0.205	0.507	0.256	0.362	0.207	0.504	0.262
		7	0.4	0%	*-0.003*	*0.232*	*-0.002*	*0.186*	*-0.003*	*0.232*	*-0.002*	*0.186*	*0.631*	*0.225*	*0.504*	*0.182*	*0.631*	*0.225*	*0.504*	*0.182*
				15%	-0.003	0.242	0.000	0.193	0.000	0.240	0.001	0.189	0.628	0.238	0.504	0.188	0.627	0.240	0.501	0.190
				30%	-0.007	0.263	-0.002	0.197	-0.013	0.254	-0.005	0.195	0.624	0.262	0.502	0.196	0.610	0.276	0.500	0.201
			0.9	0%	*-0.009*	*0.291*	*-0.011*	*0.243*	*-0.009*	*0.291*	*-0.011*	*0.243*	*0.619*	*0.281*	*0.495*	*0.239*	*0.619*	*0.281*	*0.495*	*0.239*
				15%	-0.006	0.304	-0.008	0.250	-0.012	0.300	-0.009	0.247	0.627	0.283	0.501	0.238	0.616	0.290	0.497	0.241
				30%	-0.006	0.306	-0.003	0.243	-0.005	0.304	-0.013	0.244	0.622	0.305	0.502	0.249	0.611	0.313	0.496	0.243

number indicates that the t-test comparing the group effect estimation ($\widehat{\beta group}$) and the group effect true value is significant at 5%.

§: according to Blanchin et al. [17].

Table 2: Type I error of the tests of group effect for Score Mixed model (SM) with PMS imputation and Longitudinal Rasch Mixed model (LRM) methods for different values of groups' size (N), time effect, number of items (J), latent variable correlation (ρ_θ), proportion of missing data (π) and for three cases (complete, MCAR and MNAR data). Analyses performed with an unstructured covariance matrix in SM and LRM methods.

					Complete data		MCAR		MNAR	
N	Time effect	J	ρ_θ	π	SM	LRM	SM	LRM	SM	LRM
100	0	4	0.4	0%	0.046	0.048				
				15%			0.042	0.046	0.054	0.060
				30%			0.060	0.048	0.062	0.052
			0.9	0%	0.042	0.044				
				15%			0.042	0.050	0.052	0.054
				30%			0.044	0.032	0.036	0.036
		7	0.4	0%	0.038	0.042				
				15%			0.052	0.040	0.042	0.040
				30%			0.048	0.040	0.062	0.052
			0.9	0%	0.062	0.052				
				15%			0.058	0.053	0.056	0.050
				30%			0.050	0.063	0.058	0.060
	0.2	4	0.4	0%	0.052	0.050				
				15%			0.052	0.050	0.044	0.044
				30%			0.052	0.060	0.056	0.046
			0.9	0%	0.052	0.052				
				15%			0.036	0.056	0.050	0.048
				30%			0.036	0.044	0.032	0.040
		7	0.4	0%	0.054	0.054				
				15%			0.046	0.050	0.054	0.056
				30%			0.056	0.062	0.062	0.050
			0.9	0%	0.046	0.050				
				15%			0.050	0.059	0.050	0.054
				30%			0.046	0.052	0.050	0.054
50	0	4	0.4	0%	0.048	0.052				
				15%			0.046	0.048	0.058	0.046
				30%			0.056	0.054	0.050	0.058
			0.9	0%	0.036	0.036				
				15%			0.046	0.044	0.048	0.052
				30%			0.060	0.036	0.060	0.048
		7	0.4	0%	0.074	0.072				
				15%			0.066	0.068	0.064	0.070
				30%			0.058	0.070	0.076	0.072
			0.9	0%	0.050	0.054				
				15%			0.056	0.056	0.068	0.066
				30%			0.050	0.056	0.054	0.056
	0.2	4	0.4	0%	0.046	0.050				
				15%			0.050	0.050	0.052	0.052
				30%			0.060	0.052	0.056	0.050
			0.9	0%	0.050	0.050				
				15%			0.058	0.048	0.054	0.048
				30%			0.072	0.048	0.054	0.060
		7	0.4	0%	0.072	0.070				
				15%			0.058	0.070	0.058	0.054
				30%			0.066	0.056	0.052	0.060
			0.9	0%	0.064	0.069				
				15%			0.076	0.094*	0.074	0.070
				30%			0.062	0.076	0.058	0.068

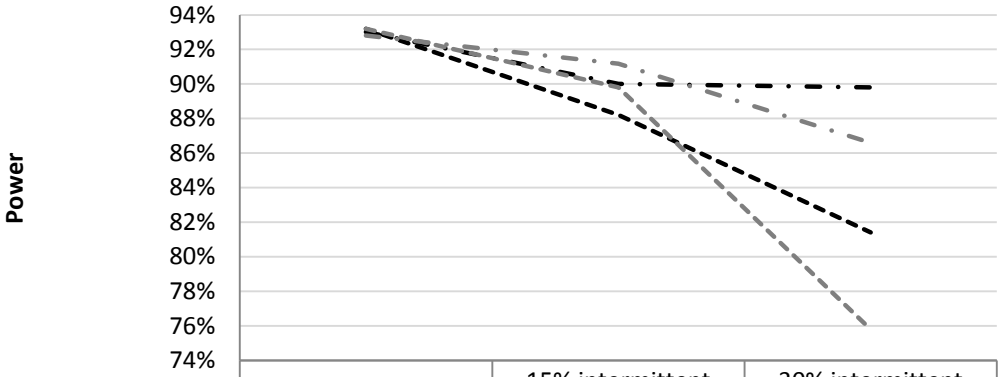
* indicates that the expected value of 5% is not included in the 95% confidence interval.
number indicates that the group effect estimation ($\widehat{\beta group}$) linked to this type I error is biased at the 5% level.

Table 3: Power of the tests of group effect for Score Mixed model (SM) with PMS imputation and Longitudinal Rasch Mixed model (LRM) methods for different values of groups' size (N), time effect, number of items (J), latent variable correlation (ρ_θ), proportion of missing data (π) and for three cases (complete, MCAR and MNAR data). Analyses performed with an unstructured covariance matrix in SM and LRM methods.

					Complete data		MCAR		MNAR	
N	Time effect	J	ρ_θ	π	SM	LRM	SM	LRM	SM	LRM
100	0	4	0.4	0%	0.932	0.928				
				15%			0.882	0.900	0.898	0.912
				30%			0.814	0.898	0.758	0.866
			0.9	0%	0.868	0.865				
				15%			0.832	0.851	0.828	0.841
				30%			0.742	0.807	0.726	0.801
		7	0.4	0%	0.972	0.968				
				15%			0.948	0.956	0.958	0.968
				30%			0.916	0.940	0.894	0.950
			0.9	0%	0.890	0.885				
				15%			0.894	0.876	0.854	0.867
				30%			0.834	0.861	0.808	0.843
	0.2	4	0.4	0%	0.948	0.950				
				15%			0.908	0.922	0.900	0.918
				30%			0.798	0.866	0.792	0.894
			0.9	0%	0.826	0.819				
				15%			0.790	0.808	0.800	0.796
				30%			0.702	0.765	0.682	0.787
		7	0.4	0%	0.970	0.970				
				15%			0.970	0.978	0.958	0.966
				30%			0.924	0.950	0.912	0.946
			0.9	0%	0.882	0.873				
				15%			0.850	0.851	0.844	0.853
				30%			0.816	0.842	0.762	0.824
50	0	4	0.4	0%	0.706	0.696				
				15%			0.616	0.643	0.642	0.674
				30%			0.510	0.588	0.502	0.612
			0.9	0%	0.584	0.568				
				15%			0.526	0.555	0.510	0.523
				30%			0.424	0.494	0.434	0.478
		7	0.4	0%	0.776	0.768				
				15%			0.738	0.744	0.742	0.730
				30%			0.696	0.732	0.640	0.688
			0.9	0%	0.710	0.696				
				15%			0.676	0.668	0.686	0.686
				30%			0.616	0.639	0.574	0.648
	0.2	4	0.4	0%	0.750	0.738				
				15%			0.678	0.690	0.656	0.706
				30%			0.566	0.654	0.496	0.631
			0.9	0%	0.582	0.573				
				15%			0.544	0.557	0.530	0.532
				30%			0.444	0.487	0.458	0.497
		7	0.4	0%	0.810	0.806				
				15%			0.770	0.791	0.758	0.786
				30%			0.696	0.729	0.664	0.720
			0.9	0%	0.606	0.598				
				15%			0.598	0.586	0.586	0.577
				30%			0.544	0.551	0.540	0.569

number indicates that the group effect estimation ($\widehat{\beta group}$) linked to this power is biased at the 5% level.

Graph 1: Comparison of power of the tests of group effect for Score Mixed model (SM) with PMS imputation and Longitudinal Rasch Mixed model (LRM) methods for one case: groups' size (N=100), no time effect, number of items (J=4), latent variable correlation (ρ_θ=0.4), proportion of missing data (π=15% or 30%) and for complete or MCAR or MNAR data. Analyses performed with an unstructured covariance matrix in SM and LRM methods.



	Complete	15% intermittent missing items	30% intermittent missing items
LRM - MCAR	93%	90%	90%
SM - MCAR	93%	88%	81%
LRM - MNAR	93%	91%	87%
SM - MNAR	93%	90%	76%

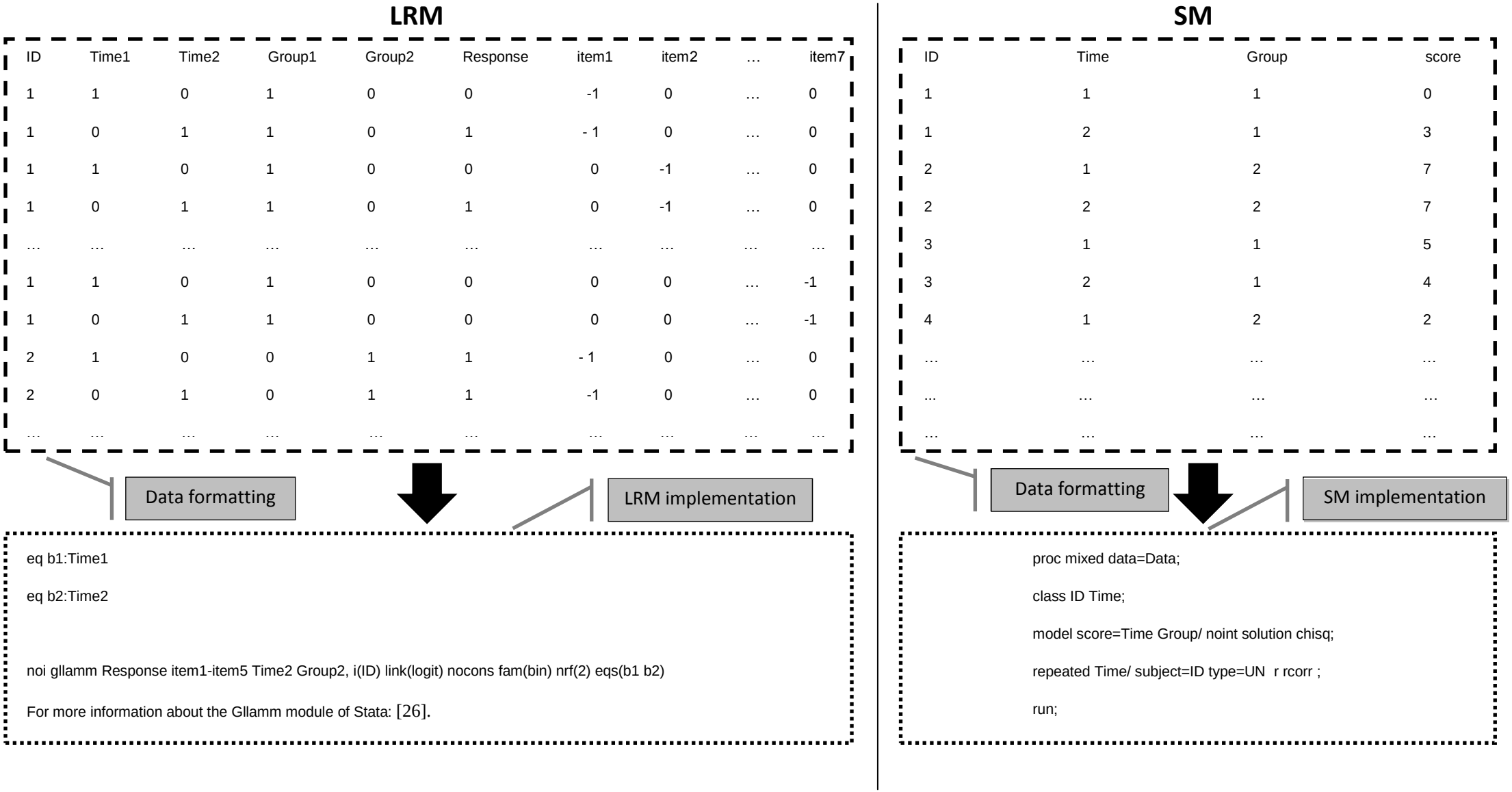


Figure 1: Example of LRM and SM implementations (with Stata and SAS, respectively) for two times of assessment, two groups and seven items with two possible responses (0 or 1).