Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2017, Vol. 26(6) 2897–2908
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280215615158
journals.sagepub.com/home/smm
SAGE

What are the appropriate methods for analyzing patient-reported outcomes in randomized trials when data are missing?

JF Hamel,[1,2] V Sebille,[1] T Le Neel,[1] G Kubis,[1] FC Boyer[3] and JB Hardouin[1]

Abstract

Subjective health measurements using Patient Reported Outcomes (PRO) are increasingly used in randomized trials, particularly for patient groups comparisons. Two main types of analytical strategies can be used for such data: Classical Test Theory (CTT) and Item Response Theory models (IRT). These two strategies display very similar characteristics when data are complete, but in the common case when data are missing, whether IRT or CTT would be the most appropriate remains unknown and was investigated using simulations. We simulated PRO data such as quality of life data. Missing responses to items were simulated as being completely random, depending on an observable covariate or on an unobserved latent trait. The considered CTT-based methods allowed comparing scores using complete-case analysis, personal mean imputations or multiple-imputations based on a two-way procedure. The IRT-based method was the Wald test on a Rasch model including a group covariate. The IRT-based method and the multiple-imputations-based method for CTT displayed the highest observed power and were the only unbiased method whatever the kind of missing data. Online software and Stata® modules compatibles with the innate mi impute suite are provided for performing such analyses. Traditional procedures (listwise deletion and personal mean imputations) should be avoided, due to inevitable problems of biases and lack of power.

Keywords

Classical test theory, item response theory, missing data, Rasch model, simulations

Subjective measurements are increasingly used in randomized clinical studies to assess patients' perception of their own health, for example quality of life (QoL), stress, or anxiety.[1,2] Such phenomena, called latent traits because they cannot be directly observed, are usually evaluated using patient reported outcomes (PRO), i.e. self-assessment questionnaires consisting of a set of items. Two strategies have been developed for analyzing PRO data: the Classical Test Theory (CTT) and the Item Response Theory (IRT). With CTT, items are combined into scores, considered as measurements of the studied latent trait.[3,4] With IRT, each of the individual item responses is modeled jointly depending on the individuals and the items characteristics, without needing to be combined into scores.[5] One of the most popular models is the Rasch model.

In randomized trials, the most appropriate methods for comparing two groups of patients on PRO measurements if data do not contain missing values are the scores comparison using t-test when using CTT, and the Wald test, performed on a Rasch model including a group covariate[6] when using IRT. These two methods are unbiased and display very similar power.[7] But when data contain (possibly informative) missing values, which method displays unbiased results is still under debate.

Missing data management differs depending on the analysis framework. With CTT, individual scores can only be calculated if data are complete. If some items responses are missing, either a listwise deletion is performed corresponding to a complete case analysis, or missing data are replaced using imputed data for score computation.

[1] EA 4275, Faculty of Pharmaceutical Sciences, University of Nantes, France
[2] Biostatistics and Methodology Unit, LUNAM Angers, CHU Angers, France
[3] Unites MPR CHU Reims, Hospital Sebastopol Reims, France

Corresponding author:
JF Hamel, EA 4275, Faculty of Pharmaceutical Sciences, University of Nantes, 1 rue Gaston Veil 44035 Nantes cedex 01, France.
Email: JeanFrancois.Hamel@chu-angers.fr

With IRT, not using an imputation method does not imply loss of information because no score calculation is required. The responses of an individual can be used even if he did not respond to all the questionnaire items. In addition, the specific objectivity property of the Rasch family models allows estimating the latent trait independently of the items used, whether responses are observed to all items or not, without any imputation being required.[5,8] When an individual completes a questionnaire, his (unobservable) latent trait is the same for all the items. If the data fit a Rasch family model (characterized by the specific objectivity property), an unbiased estimate of the latent trait can be obtained on a subset of the items, regardless of the item selection process, if at least an item has been answered.

The aim of our study was to compare two different group-comparison methods when data contain possibly informative missing values: a CTT-based method (scores comparison using a t-test) and an IRT-based method (the Wald-test performed using random effects Rasch model including a group covariate).[7]

1 Methods

1.1 Simulation study

In our study, the empirical properties of the studied statistical methodologies were explored using Monte-Carlo simulations. Data were simulated to represent situations encountered in real-life studies such as a randomized trial whose aim would be to assess the impact of a new treatment on QoL. Patients could be characterized by both their age and their quality of life, and may not have answered all of the PRO's items. Several reasons could account for an item non-response. First, the item non-response could be completely random, i.e. no covariates could explain this non-response. Second, an observable patient's characteristic as the age of respondents could explain the random process: the elderly presenting a higher non-response rate.[9] Third, the latent trait of interest may be the cause of an item non-response: patients with poorer QoL presenting a higher non-response rate.[10] Such data were simulated as follows: two samples of equal size *A* and *B*, corresponding to the treatment groups to be compared, were generated. Each group was divided into two (observable) subgroups *.1* and *.2* of equal size (respectively, *A1* and *A2*, and *B1* and *B2*), corresponding to younger and older patients, respectively. A latent trait value was simulated for each simulated individual, representing his/her QoL level.

The simulated PRO were composed of dichotomous items. The individual responses probabilities were computed using a Rasch model. Hence, the simulated PRO were assumed to have been previously validated both with IRT and with CTT, as all the assumptions underlying the CTT are included in those underlying the IRT.[11]

Missing data were simulated as a function of the average probability of an item non-response: P_{mean}, the difference in non-response probability related to the sub-group membership: P_{gp}, and the magnitude of the variation of the non-response probability related to the individual latent trait: P_{theta}. When P_{gp} and P_{theta} were set to 0 (i.e. the non-response rate was neither influenced by the respondents' age nor by their QoL), missing data were completely random. When P_{gp} was set different from 0 and P_{theta} equal to 0, missing data depended only on an observable covariate. Finally, when P_{theta} was set different from 0, missing data depended on the unobservable latent trait. The individual non-response probability was calculated as follows:

$$P(R_i = r_i|\theta_i, \delta_i, H_i = h_i) = P_{mean} + h_i\frac{P_{gp}}{2} + P_{theta}\left(\frac{1}{1+\exp\theta_i} - \frac{1}{2}\right)$$

where R_i is a dummy random variable coded 1 or 0 representing, respectively, the presence or the absence of a response to the item *j* by the individual *i* and H_i is a dummy variable representing the sub-group membership of individual i (coded -1 if the *i*th individual is part of the sub-group *.1*, and 1 if not). The non-response probability depending on the simulation parameters is illustrated in Figure 1.

The simulation parameters used to simulate such data are presented in Table 1. The total number of simulation parameters combinations was 2304. Each of them was replicated 1000 times.

1.2 Statistical analysis

1.2.1 The score analysis

The individual scores were defined as the number of items with positive responses. The average scores of each group were then compared with a t-test. Three distinct strategies were explored for the missing data management:

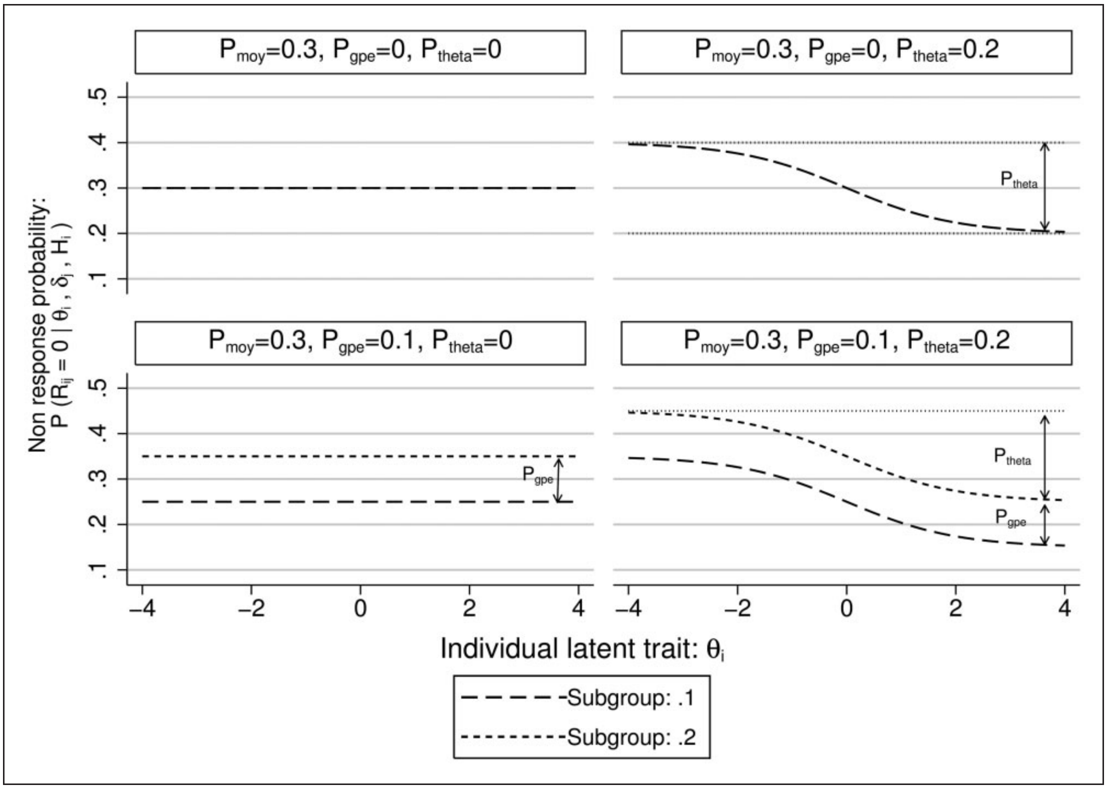

Figure 1. Item non-response probability according to the individual latent trait, the subgroup membership and the non-response simulation parameters. P_{mean}: average probability of an item non-response, P_{gp}: maximum variation of probability related to the observed group membership, P_{theta}: maximum variation of probability related to the individual latent trait.

- Only complete questionnaires were analyzed (complete case analysis: "score$_{cc}$" method).
- Missing responses were imputed by personal mean scores, consisting for a patient with less than half of missing responses in imputing a missing value by his observed mean response ("score$_{mean\text{-}imp}$" method). Such a method is for example recommended in the manual of several widely used questionnaires such as the SF-36 or the QLQ-C30.[12,13]
- Missing responses were imputed using multiple imputations based on the Two-Way methodology,[14–16] the number of imputations being set to 10 ("score$_{Two\text{-}Way}$" method). The two-way method consists in modeling items responses using a two-way ANOVA: the observed response of the ith individual to the item j: x_{ij} is defined as a function of an individual effect: a_i (considered as a random effect normally distributed), an item effect: β_j (considered as a fixed effect) and an error effect: ε_{ij} normally distributed with mean equal to 0: $X_{ij} = a_i + \beta_j + \varepsilon_{ij}$. When data are missing, the parameters of this model are estimated using an EM algorithm, and finally used for imputing missing data, taking into account both the individuals and the items characteristics.[17]

1.2.2 The latent trait analysis

The latent trait analysis was performed using a random effects Rasch model including a group covariate, and the significance of this covariate was tested using a Wald test ("IRT$_{cov}$" method).[6,7] Such a Rasch model is detailed below. Let g_i be a dummy variable characterizing the ith individual ($g_i = \{0;\ 1\}$), X_{ij} the dichotomous variable representing the response of this individual to the jth item ($x_{ij} = 0$ for a negative response and $x_{ij} = 1$ for a positive response), θ his residual latent trait (drawn from a normal distribution with mean equal to 0) and δ_j the difficulty of item j. The random effects Rasch model with a group covariate can then be written as follows

$$P(X_{ij} - x_{ij}\theta, g_i, \gamma, \delta_j) = \frac{\exp(x_{ij}(\theta + \gamma.g_i - \delta_j))}{1 + \exp(\theta + \gamma.g_i - \delta_j)}.$$

Table 1. Possible values of the different simulation parameters.

Parameters	Values
Sample size: $n = n_A + n_B$	100 200 400 800
Differences between the latent traits means: γ	0 0.2σ 0.5σ 0.8σ
Latent trait distribution	• Normally distributed ○ Mean: $\mu_A = \frac{-\gamma}{2}$ $\mu_B = \frac{\gamma}{2}$ ○ Variance: $\sigma^2 = \sigma_A^2 = \sigma_B^2 = 1$
Number of items: j	5 10
Items difficulties distribution	• Percentiles of a standardized normal distribution • Percentiles of an equiprobable mixture of two Gaussian distributions with parameters (**M**; **Σ**) ○ $M = \begin{pmatrix} -\sigma \\ \sigma \end{pmatrix}$ ○ $1\Sigma = \begin{pmatrix} (0.3\,\sigma)^2 & 0 \\ 0 & \sigma^{\prime}2 \end{pmatrix}$
Average probability of an item non-response: P_{mean}	0% 10% 20% 30%
Difference in non-response probability related to the sub-group membership: P_{gp}	0% 10% 20% 30%
Magnitude of the non-response probability variation related to the individual latent trait: P_{theta}	0% 10% 20% 30%

1.2.3 Comparison of methods

Four criteria were studied: the position bias, the dispersion bias, the type I error, and the power.

For each combination of simulation parameters:

- When the methodology was based on IRT, a position bias was defined as a difference different from 0 between the observed and the simulated latent trait difference between groups. A dispersion bias was defined as a difference between the observed and the simulated latent trait variance different from 0.
  - The latent trait difference between groups was estimated by the average of the observed group effect estimates over the 1000 replicated simulations. A position bias of less than 0.1σ was considered as not relevant in practice.
  - The variances of the two groups were assumed equal: $\sigma_A^2 = \sigma_B^2 = \sigma^2$. They were estimated by the average of the observed residual latent traits variances σ_{Res}^2 over the 1000 replicated simulations: σ_{Obs}^2. A bias of less than $0.1\sigma^2$ was considered as not relevant in practice.
- When the methodology was based on CTT, a position bias was defined as a difference different from 0 between the observed and the simulated score difference between groups. A dispersion bias was defined as a difference between the observed and the simulated score variances different from 0.
  - The scores variances of the two groups were assumed equal, and were estimated by the average of the observed scores variances over the 1000 replicated simulations. A bias of less than $0.1\sigma_S^2$ was considered as not relevant in practice.
  - The score difference between groups was estimated by the average of the differences between the observed means of the scores of the groups A and B over the 1000 replicated simulations. A position bias of less than $0.1\sigma_S$ was considered as not relevant in practice.
  - We detail in the Appendix how the values of the simulated scores differences and variances were computed using the simulation parameters.
- The type-I error was obtained by calculating the proportion of rejection of the null hypothesis among the 1000 replications of each parameters combinations with γ set to 0, and compared to the expected rejection proportion (0.05) using a chi-square test.
- Power was obtained depending on the parameters combinations by calculating the proportion of rejection of the null hypothesis among the 1000 replications of each parameters combination with γ set different from 0. A power variation of less than 0.05 was considered as not relevant in practice.

The effect of the simulation parameters on the observed biases and powers was studied using linear models for evaluating separately the effects of each parameter and their potential interactions. The only considered interactions were those with effects relevant in practice for at least one of the considered methods. For the analysis, P_{mean}, P_{gp} and P_{theta} were considered as continuous variables whereas the simulated difference, the number of items and the item difficulties distribution were considered as qualitative variables. The validity of the different models was checked by studying residual plots.

Simulations and statistical analyses were performed with the Stata 12.1 software and the Gllamm package.[18,19]

2 Results

2.1 Position Biases

When data were complete, none of the comparison methods led to any position bias that was relevant in practice (Table 2).

When data were missing, positions biases were observed for the "score$_{cc}$" method when missing data depended on the unobservable latent trait, and for the "score$_{mean\text{-}imp}$" for all types of missing data. These position biases resulted systematically in an underestimation of the group scores difference. For the "score$_{cc}$" method, increasing P_{theta} led to position biases that were more pronounced when the simulated difference γ and the number of items both increased. For the "score$_{mean\text{-}imp}$" method, increasing the number of items, P_{mean} or P_{theta} led to position biases that were more pronounced for a large simulated difference γ (Figure 2). Increasing P_{mean} led to more pronounced position biases when the number of items was large.

The "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" method did not led to any position biases whatever the type of missing data. The sample size and the items difficulty distribution did not affect the position biases.

Table 2. Effects of the simulation parameters on the observed position biases for the different methodologies for comparing groups on subjective measurements estimated using linear regression.

Parameter	Score$_{cc}$	Score$_{mean\text{-}imp}$	Score$_{two\text{-}way}$	IRT$_{cov}$
n (+100)	−0.001	0.000	0.000	0.000
$\gamma=0$	0	0	0	0
$\gamma=0.2$	0.006	0.011	−0.001	0.005
$\gamma=0.5$	0.029	0.037	−0.003	0.020
$\gamma=0.8$	0.051	0.056	−0.004	0.031
$j=10 \mid \gamma=0$	−0.006	−0.002	−0.001	−0.001
$j=10 \mid \gamma=0.2$	−0.015	−0.038	0.000	−0.004
$j=10 \mid \gamma=0.5$	−0.042	−0.076	0.001	−0.002
$j=10 \mid \gamma=0.8$	−0.079	**−0.127**	−0.002	−0.004
D_{diff} = *Norm.*	0	0	0	0
D_{diff} = *Mixt.*	−0.004	0.004	0.001	−0.007
P_{mean} (+30%) $\mid \gamma=0$	−0.014	−0.007	0.002	0.000
P_{mean} (+30%) $\mid \gamma=0.2$	−0.011	−0.076	0.003	0.003
P_{mean} (+30%) $\mid \gamma=0.5$	−0.005	**−0.185**	−0.002	0.008
P_{mean} (+30%) $\mid \gamma=0.8$	−0.016	**−0.288**	0.001	0.011
P_{gpe} (+30%)	0.013	0.010	0.001	0.001
P_{theta} (+30%) $\mid \gamma=0$	0.006	0.002	−0.005	0.000
P_{theta} (+30%) $\mid \gamma=0.2$	−0.030	0.033	−0.004	0.000
P_{theta} (+30%) $\mid \gamma=0.5$	**−0.098**	0.082	0.004	−0.002
P_{theta} (+30%) $\mid \gamma=0.8$	**−0.148**	**0.130**	0.001	−0.003
Cons.	0.009	0.000	0.001	0.005

Note: The reported values correspond to the parameters associated with each of the factors that may influence the observed position bias, estimated using linear regression.

D_{diff}: items difficulties distribution, *j*: number of items, *n*: sample size, γ: simulated difference, P_{mean}: average probability of an item non-response, P_{gp}: maximum variation of probability related to the observed group membership, P_{theta}: maximum variation of probability related to the individual latent trait, *Norm.*: normal distribution, *Mixt.*: equiprobable mixture of two normal distributions, *Cons.*: constant. P_{gp} and P_{theta} set at 0 corresponds to completely random item non-response. P_{gp} set higher than 0 corresponds to observable covariate dependent missing data. P_{theta} set higher than 0 corresponds to latent trait dependent missing data. Methodologies with relevant observed position bias variations appear in bold.

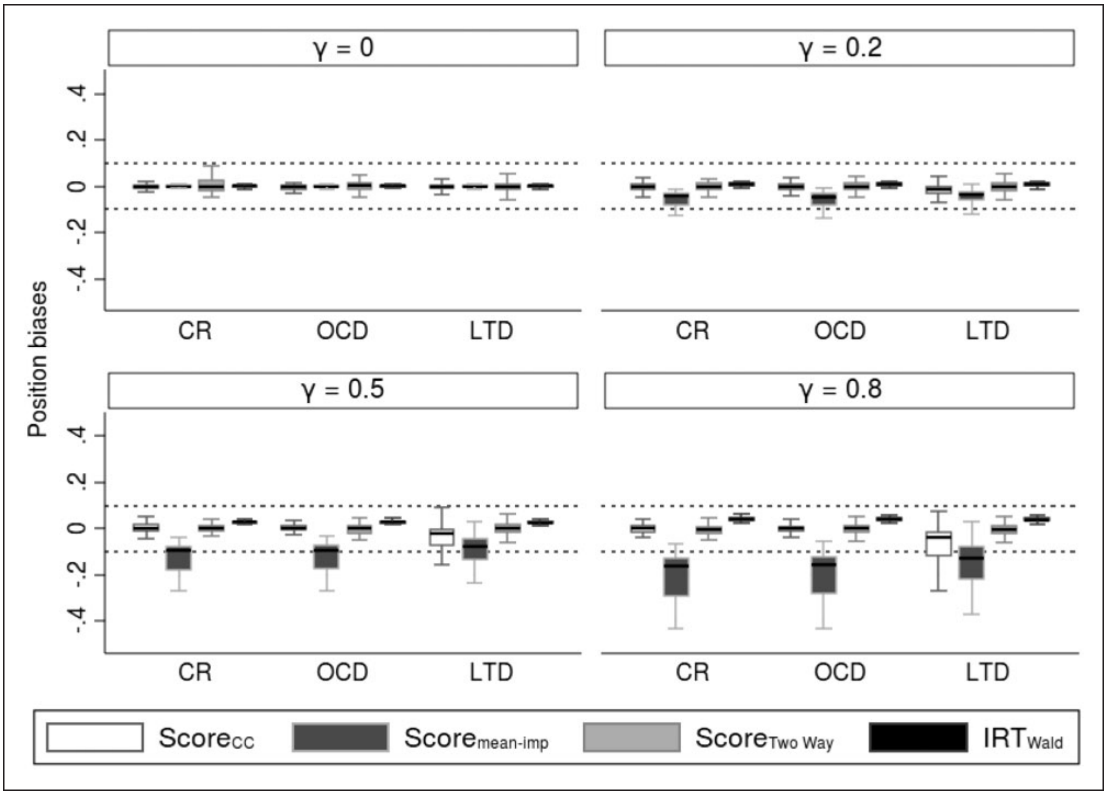

Figure 2. Observed position biases for the different methodologies according to the simulated difference and the type of missing data. CR: completely random missing data, OCD: observable covariate-dependent missing data, LTD: latent trait-dependent missing data.

2.2 Dispersion biases

When data were complete, none of the comparison methods led to any dispersion bias that was relevant in practice (Table 3). An underestimation of the latent trait variance was observed for the "score$_{cc}$" method when missing data depended on an unobservable latent trait and for the "score$_{mean\text{-}imp}$" when data were missing whatever the type of missing data. An overestimation of the latent trait variance was observed for the "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" methods when data were missing whatever the type of missing data.

For the "score$_{cc}$" method, increasing P_{theta} led to dispersion biases when the number of items was large. For the "score$_{mean\text{-}imp}$" method, increasing P_{gp} or P_{theta} led to dispersion biases that were more marked when the number of items was large (Figure 3).

The sample size and the simulated difference did not seem to impact the dispersion biases.

2.3 Type I error and power

The type I errors were not significantly different from 0.05 for the "score$_{cc}$", "score$_{two\text{-}way}$" and "IRT$_{cov}$" methods. The "score$_{mean\text{-}imp}$" tended to slightly minimize the type I error (average observed "score$_{mean\text{-}imp}$" type-I error = 4.51%). When data were complete, all the methods led to similar power. When data were missing, the "IRT$_{cov}$" and "score$_{two\text{-}way}$" methods led to the highest observed powers and the "score$_{cc}$" method to the lowest observed power (Table 4).

Increasing the sample size, the number of items and the simulated difference was associated with an increase of the observed power (Figure 4). The difficulties distribution did not affect the observed power.

Increasing P_{mean} resulted in a decrease of the observed power. This power decrease was more pronounced for the "score$_{cc}$" method and more moderate for the "IRT$_{cov}$" and "score$_{two\text{-}way}$" methods.

The type of missing data did not impact the observed power for any of the studied methods. For the "score$_{cc}$" method, increasing P_{gp} resulted in an increase of the observed power.

Table 3. Effects of the simulation parameters on the observed dispersion biases for the different methodologies for comparing groups on subjective measurements estimated using linear regression.

Parameter	$Score_{cc}$	$Score_{mean\text{-}imp}$	$Score_{two\text{-}way}$	IRT_{cov}
n (+100)	0.000	0.000	0.004	0.002
$\gamma=0$	0	0	0	0
$\gamma=0.2$	0.004	0.000	0.000	−0.001
$\gamma=0.5$	0.014	0.007	0.001	0.001
$\gamma=0.8$	0.015	0.024	0.001	−0.003
$j=10 \mid D_{diff}=Norm.$	**0.097**	−0.023	0.031	−0.032
$j=10 \mid D_{diff}=Mixt.$	0.076	**0.111**	0.023	−0.035
$D_{diff}=Norm.$	0	0	0	0
$D_{diff}=Mixt.$	−0.002	0.057	0.019	−0.029
P_{mean} (+30%) $\mid j=5$	−0.011	**−0.429**	**0.386**	**0.121**
P_{mean} (+30%) $\mid j=10$	**−0.105**	**−1.727**	**0.214**	0.066
P_{gp} (+30%) $\mid j=5$	0.005	**0.092**	0.048	0.000
P_{gp} (+30%) $\mid j=10$	0.072	**0.438**	0.031	0.002
P_{theta} (+30%) $\mid j=5$	−0.077	**0.160**	0.004	−0.005
P_{theta} (+30%) $\mid j=10$	**−0.776**	**0.778**	0.004	−0.005
Cons.	0.002	−0.015	−0.101	0.132

Note: The reported values correspond to the parameters associated with each of the factors that may influence the observed dispersion bias, estimated using linear regression. D_{diff}: items difficulties distribution, j: number of items, n: sample size, γ: simulated difference, P_{mean}: average probability of an item non-response, P_{gp}: maximum variation of probability related to the observed group membership, P_{theta}: maximum variation of probability related to the individual latent trait, *Norm.*: normal distribution, *Mixt.*: equiprobable mixture of two normal distributions, *Cons.*: constant. P_{gp} and P_{theta} set at 0 corresponds to completely random item non-response. P_{gp} set higher than 0 corresponds to observable covariate dependent missing data. P_{theta} set higher than 0 corresponds to latent trait dependent missing data. Methodologies with relevant observed dispersion bias variations appear in bold.

3 Example

We illustrate the results of this simulation study using data coming from a pilot study used to compare the pain level of patients suffering from muscular dystrophies, depending on the kind of muscular dystrophies. The ethics committee of Reims, France granted approval for the study. Patients were recruited from the University Hospital of Reims as follows: 52 patients with Steinert's disease and 95 patients with others muscular dystrophies. QoL was evaluated using the Nottingam Health Profil (NHP) questionnaire. The main outcome was the score on the pain sub-scale, the score being computed as the weighted sum of the items according to the NHP manual.[20]

First, only the questionnaires without any missing data were used to calculate the score ("$score_{cc}$" method). Forty-seven of the 52 (90%) patients with Steinert disease, and 72 patients of the 95 (76%) with other muscular dystrophies had responded to all of these sub-scale items. The mean scores for the pain sub-scale in each group were 35.4 (SD=28.2) for patients with a Steinert's disease and 28.0 (SD=27.9) for patients with other diseases. The difference between these mean scores was not statistically significant ($p=0.161$).

Next, missing responses were imputed by the average observed responses for each individual who responded to at least half of the items ("$score_{mean\text{-}imp}$" method). Also, 51 of the 52 (98%) patients with Steinert disease, and all of the patients with others muscular dystrophies had responded to at least half of these sub-scale items. The mean scores for the pain sub-scale in each group were then 35.3 (SD=27.5) for patients with a Steinert's disease and 27.0 (SD=28.4) for patients with other diseases. The difference between these mean scores was still not statistically significant ($p=0.091$).

Then, the missing responses were imputed using the Two-Way methodology ("$score_{Two\text{-}Way}$" method). All of the patients could be included. The mean scores for the pain sub-scale in each group were 35.7 (SD=27.2) for patients with a Steinert's disease and 27.2 (SD=28.4) for patients with other diseases. The difference between these mean scores was not statistically significant ($p=0.079$).

Finally, a random effect Rasch model including a group effect was fitted on these data. The global fit of the Rasch model was not rejected by the R1m test ($p=0.329$).[21] The group-comparison was performed by testing the nullity of the parameter associated with the group-covariate ("IRT_{cov}" method). The estimation of the difference between the mean levels of the latent trait of the two groups of patients was 0.649 and the variance of the latent

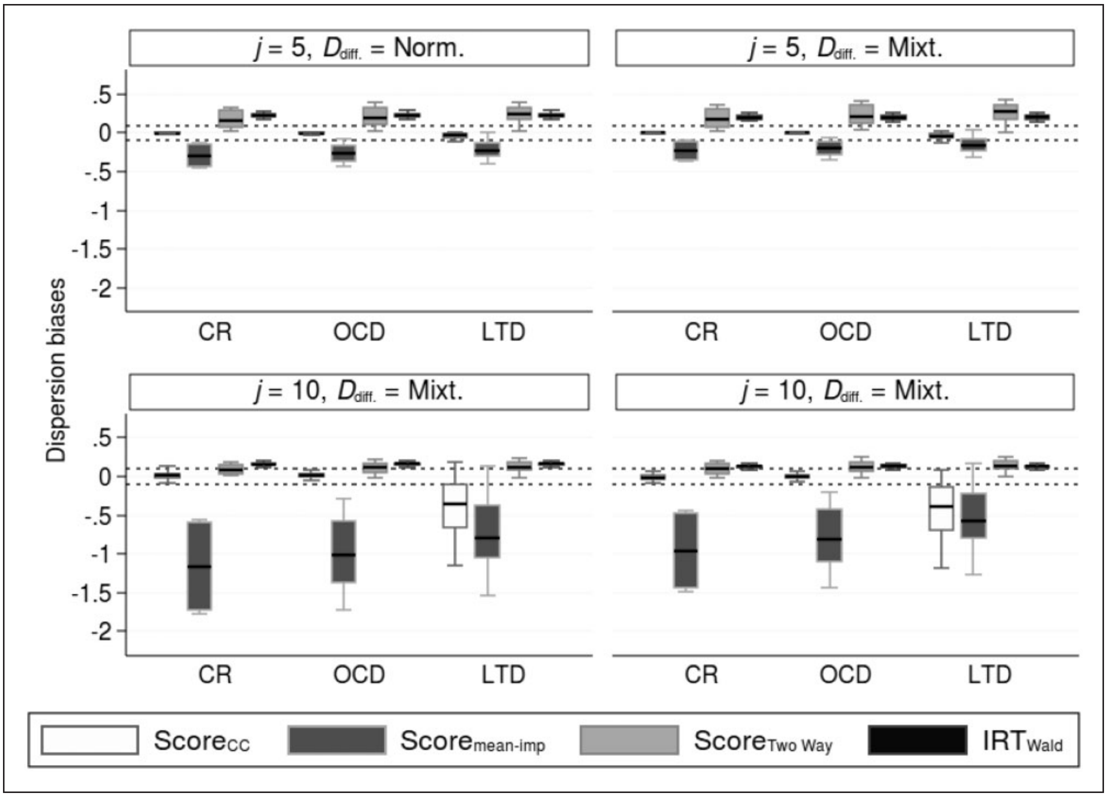

Figure 3. Observed dispersion biases for the different methodologies according to the number of items, the items difficulties distribution and the type of missing data. CR: completely random missing data, OCD: observable covariate dependent missing data, LTD: latent trait-dependent missing data.

trait was 3.84. The mean levels of the two groups latent trait were statistically significantly different ($p = 0.044$). With this method, we could conclude that the pain of patients suffering from Steinert's disease was higher than the pain of patients suffering from others muscular dystrophies.

4 Discussion

When data were missing, the type of missing data (completely random, dependent on an observable covariate or on the unobservable latent trait) had no effect on the observed properties of the "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" methods. Position biases were only observed when missing data depended on the latent trait for the "score$_{cc}$" method, whereas the "score$_{mean\text{-}imp}$" method displayed position biases whatever the type of missing data. These differences of performance could be explained by the type of missing data and the chosen comparison methodology.

The "score$_{mean\text{-}imp}$" imputation method is still the most widely used for PRO analysis.[22] However, data imputed with this method are known to be biased.[8,23] Such biases are then directly related to the imputation method, and not to the type of missing data. As the number of imputed responses increases with the number of items for a given non-response rate, the importance of biases also increases with the number of items.[21] Such an imputation method should not be recommended for PRO analysis.

When the missing data process is completely random, a listwise deletion process should not produce any bias but a decrease of power, by decreasing the sample size,[24] and this is what we actually observed using the "score$_{cc}$" method. When missing data depended on an observable covariate, a listwise deletion process should result in biases as the complete cases cannot be considered as a random sample of all the cases. Surprisingly, we did not observe such a phenomenon: the position biases were still irrelevant for the "score$_{cc}$" method when missing data depended on an observable covariate. Finally, when missing data depended on the unobserved latent trait, the "score$_{cc}$" method led to relevant biases.

The "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" methods did not lead to any position bias whatever the simulation parameters combination, the non-response rate and the type of missing data. For the "IRT$_{cov}$" method, this

Table 4. Effects of the simulation parameters on the observed powers for the different methodologies for comparing groups on subjective measurements estimated using linear regression.

Parameter	Score$_{cc}$	Score$_{mean\text{-}imp}$	Score$_{two\text{-}way}$	IRT$_{cov}$
$\gamma = 0.2$	0	0	0	0
$\gamma = 0.5$	**0.352**	**0.400**	**0.407**	**0.435**
$\gamma = 0.8$	**0.540**	**0.770**	**0.780**	**0.806**
n (+100) $\mid \gamma = 0.2$	0.017	**0.052**	**0.054**	**0.056**
n (+100) $\mid \gamma = 0.5$	**0.066**	**0.080**	**0.077**	**0.073**
n (+100) $\mid \gamma = 0.8$	**0.075**	0.029	0.025	0.022
$j = 10 \mid \gamma = 0.2$	−0.007	**0.052**	**0.058**	**0.054**
$j = 10 \mid \gamma = 0.5$	−0.007	**0.066**	**0.069**	**0.073**
$j = 10 \mid \gamma = 0.8$	**0.068**	0.027	0.029	0.030
D_{diff} = *Norm.*	0	0	0	0
D_{diff} = *Mixt.*	−0.008	−0.013	−0.005	−0.010
$P_{mean} \mid \gamma = 0.2 \mid j = 5$	**−0.153**	**−0.080**	**−0.073**	−0.045
$P_{mean} \mid \gamma = 0.5 \mid j = 5$	**−0.529**	**−0.133**	**−0.097**	**−0.067**
$P_{mean} \mid \gamma = 0.8 \mid j = 5$	**−0.514**	**−0.087**	**−0.055**	−0.038
$P_{mean} \mid \gamma = 0.2 \mid j = 10$	**−0.178**	**−0.077**	**−0.057**	−0.035
$P_{mean} \mid \gamma = 0.5 \mid j = 10$	**−0.696**	**−0.094**	**−0.053**	−0.043
$P_{mean} \mid \gamma = 0.8 \mid j = 10$	**−0.869**	**−0.051**	−0.021	−0.018
P_{theta} (+30%)	−0.002	0.008	0.001	0.000
P_{gp} (+30%) $\mid \gamma = 0.2$	0.026	−0.019	0.000	−0.001
P_{gp} (+30%) $\mid \gamma = 0.5$	**0.133**	−0.021	−0.013	−0.001
P_{gp} (+30%) $\mid \gamma = 0.8$	**0.187**	−0.009	−0.001	−0.002
Cons	0.167	0.081	0.070	0.061

Note: The reported values correspond to the parameters associated with each of the factors that may influence the observed power, estimated using linear regression. D_{diff}: items difficulties distribution, j: number of items, n: sample size, γ: simulated difference, P_{mean}: average probability of an item non-response, P_{gp}: maximum variation of probability related to the observed group membership, P_{theta}: maximum variation of probability related to the individual latent trait, *Norm.*: normal distribution, *Mixt.*: equiprobable mixture of two normal distributions, *Cons.*: constant. P_{gp} and P_{theta} set at 0 corresponds to completely random item non-response. P_{gp} set higher than 0 corresponds to observable covariate dependent missing data. P_{theta} set higher than 0 corresponds to latent trait dependent missing data. Methodologies with relevant observed power variations appear in bold.

could be explained by the specific objectivity property of the Rasch model: if the data fit a Rasch family model, unbiased estimates of the latent trait can be obtained even if they are only based on a single filled item, regardless of the number of missing items and the reasons for which these items were not completed. However, the accuracy of this latent trait estimate will be closely linked to the number of missing items, explaining both the overestimation of the latent trait variance and the power decrease observed with the "IRT$_{cov}$" method when the non-response rate increased.

The "score$_{Two\text{-}Way}$" method, based on a two-way ANOVA, is especially suitable for imputations of PRO missing responses, since it allows performing imputations by taking into account both the individual and the item characteristics: individual characteristics are modeled using ANOVA by estimating the parameters associated with the individual effects, and the item characteristics by the parameters associated with the items. The parameters associated with the individual effects can be understood as manifest variables illustrating the individual latent traits. Then, consistent values can be imputed even if the missing data process depends on the latent trait of interest, because the unobserved latent trait is taken into account through the ANOVA individual effects when performing imputations with the "score$_{Two\text{-}Way}$" method.

The "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" methods were the most powerful methods whatever the type of missing data. The "score$_{mean\text{-}imp}$" method led to slightly lower power than those observed with the "score$_{Two\text{-}Way}$" and "IRT$_{cov}$" methods, despite a significant underestimation of the mean score differences between groups. This phenomenon was due to the simultaneous increase of the dispersion biase. The "score$_{cc}$" method was clearly the less powerful method, because of the listwise deletion process resulting in a drastic reduction of sample-size.

The power rise associated with the number of items increase could be linked to the subjectivenature of the latent traits: since latent variables are not directly observable, their estimate accuracy is largely dependent on the tool used to perform these estimations. Increasing the number of items of a questionnaire leads to an increase of the reliability of the latent traits estimation, and thus to an increase of the tests' power.[25]

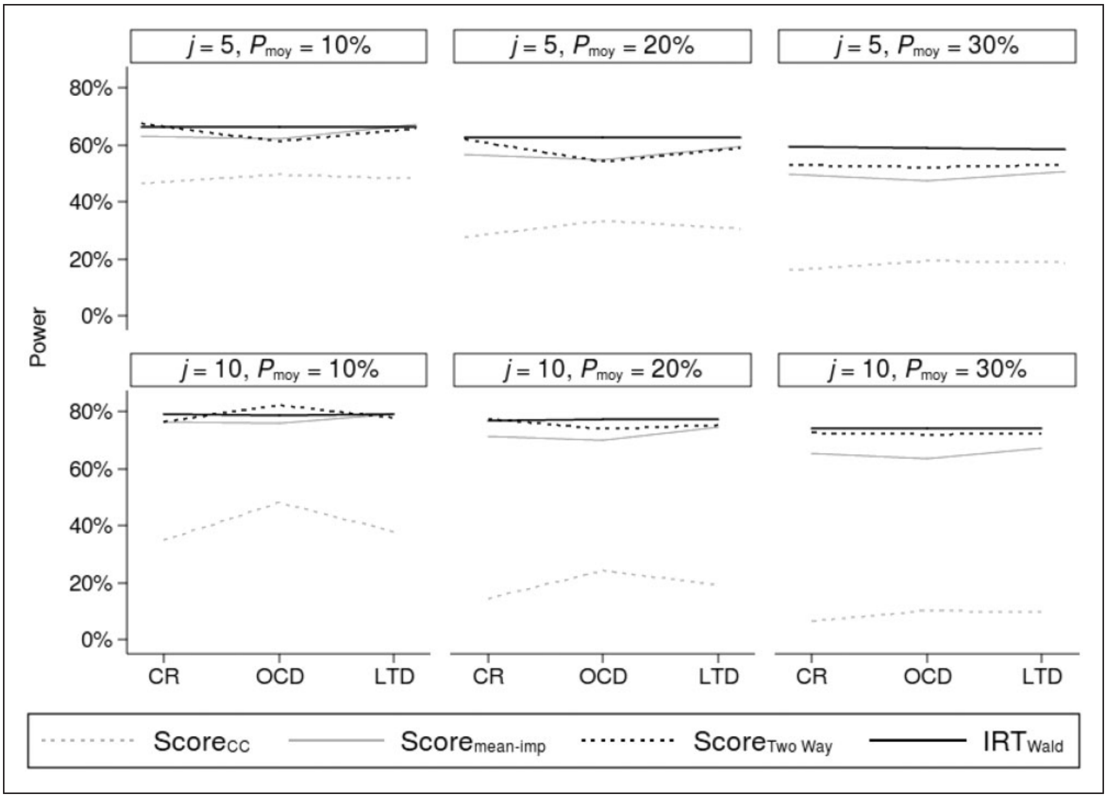

Figure 4. Observed powers for the different methodologies according to the average probability of an item non response, the number of items and the type of missing data. γ is set at 0.5, n is set at 200 and the items' difficulties are normally distributed. CR: completely random missing data, OCD: observable covariate dependent missing data, LTD: latent trait-dependent missing data.

We developed several modules available online for the Stata® statistical software: the mi_twoway Stata® module (based on the Stata® mi impute suite and fully compatible with the innate mi estimate commands) allowing performing PRO analyses based on the "score$_{Two\text{-}Way}$" method, and the pcmodel Stata® module allowing performing such analyses based on the "IRT$_{cov}$" method. Moreover, the free PRO-online software is proposed for performing analyzes using Rasch models (http://pro-online.univ-nantes.fr). We believe that such programs may facilitate the use of these methods by researchers.

Some limitations should be recorded. First, we only considered Gaussian distributions of the latent traits. In real life situations, the latent trait could be not normally distributed but follow a bounded and asymmetrical distribution. Indeed, if one considers for example pain as the studied latent variable, one should consider a lower bound distribution, corresponding to a lack of pain. A log-normal distribution could then be considered for example. Second, biases detection was carried out by using thresholds corresponding to the minimum biases considered as relevant in practice. A different choice of thresholds could lead to different results. However, the graphical representation of the biases according to different types of missing data allowed assessing the biases importance independently of the selected thresholds. Third, we only considered questionnaires assumed to be validated both with IRT and CTT. We did not consider the case of questionnaires validated with CTT but not with IRT for instance, although it can be encountered in real life studies. Finally, Rasch models, requiring necessarily dichotomous items, are models that may seem too restrictive to be applied in real situations. It appears necessary to pursue this study by analyzing extensions of the Rasch model, allowing for polytomous items analysis, such as the Partial Credit Model.

5 Conclusion

When data are missing, standard analysis procedures such as complete case analyses or personal mean imputations are not appropriate for comparing two groups of patients on PRO measurements. Specific imputation methods as the two-way procedure should be considered when performing such analyses within the CTT framework. IRT-based methods, such as using a Wald test performed on a random effects Rasch model including a group covariate, may also be appropriate.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Lipscomb J, Gotay CC and Snyder CF. Patient-reported outcomes in cancer: a review of recent research and policy initiatives. *CA: Cancer J Clinician* 2007; **57**: 278–300.
2. Willke JR, Burke LB and Erickson P. Measuring treatment impact: A review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Control Clin Trial* 2004; **25**: 535–552.
3. Walters SJ, Campbell MJ and Lall R. Design and analysis of trials with quality of life as an outcome: A practical guide. *J Biopharm Stat* 2001; **11**: 155–176.
4. Hambleton RK and Jones RW. Comparison of classical test theory and item response. *Educ Measure: Issues Practice* 1993; **12**: 38–47.
5. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press, 1960.
6. Zwinderman AH. A generalized rasch model for manifest predictors. *Psychometrika* 1991; **56**: 589–600.
7. Hamel JF, Hardouin JB, Le Neel T, et al. Biases and power for groups comparison on subjective health measurements. *PloS ONE* 2012; **7**: e44695.
8. Hardouin JB, Conroy R and Sébille V. Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Med Res Methodol* 2011; **14**: 11–105.
9. Matsuda T, Marche H, Grosclaude P, et al. Participation behavior of bladder cancer survivors in a medical follow-up survey on quality of life in France. *Eur J Epidemiol* 2004; **19**: 313–321.
10. Hensing TA, Peterman AH, Schell MJ, et al. The impact of age on toxicity, response rate, quality of life, and survival in patients with advanced, stage IIIB or IV nonsmall cell lung carcinoma treated with carboplatin and paclitaxel. *Am Cancer Soc* 2003; **98**: 779–788.
11. Holland PW. Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika* 2003; **68**: 123–149.
12. Aaronson NK and Ahmedzai S. Bergman Bea: The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Instit* 1993; **85**: 365–376.
13. Ware JE and Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473–483.
14. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.
15. Van Ginkel JR, Van der Ark LA, Sijtsma K, et al. Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computat Stat Data Analysis* 2007; **51**: 4013–4027.
16. Bernaards CA and Sijtsma K. Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behav Res* 2000; **35**: 321–364.
17. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.
18. Rabe-Hesketh S, Pickles A and Taylor C. Generalised, linear, latent and mixed models. *Stata Tech Bull* 2000; **53**: 47–57.
19. Zheng X and Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. *The Stata J* 2007; **7**: 313–333.
20. Hunt SM, McKenna SP, MCEwen J, et al. The Nottingham Health Profile: subjective health status and medical consultations. *Social Sci Med* 1981; **15**: 221–229.
21. Glas CAW. The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* 1988; **53**: 525–546.
22. Dondersa ART, van der Heijdenc GJMG, Stijnend T, et al. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; **59**: 1087–1091.
23. White IR and John B. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010; **29**: 2920–2931.
24. Little RJA and Rubin DB. *Statistical analysis with missing data*. Wiley: New York, 1987.

25. Sébille V, Hardouin JB, Le Néel T, et al. Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients – A simulated study. *BMC Med Res Methodol* 2010; **25**: 10–24.

Appendix 1

The score differences between groups were not defined by the simulation parameters, as opposed to the latent trait differences. Hence the simulation parameters did not directly allow estimating position biases for the methods based on the score (defined as the differences between the observed and the simulated score differences). The same, the score variances were not defined by the simulation parameters, whereas the latent trait variances were. Dispersion biases (defined as the differences between the observed and the simulated score variances) were not directly estimable based on the simulation parameters. However, it was possible to calculate both the simulated score differences between groups and score variances as follows:

The true value of score group effect γ_S was approached by the difference of the expected score in each group.

$$\gamma_S = E(S_i|\,g = B) - E(S_i|\,g = A)$$

The expected score in each group was computed as follows

$$\begin{aligned} E(S_i|\,g) &= E\left(\sum_j x_{ij}|\,g\right) \\ &= \sum_j E\left(x_{ij}|\,g\right) \\ &= \sum_j P\left(x_{ij} = 1|\,g\right) \\ &= \sum_j \int_{-\infty}^{+\infty} P\left(x_{ij} = 1\right)\Phi\left(\theta|\mu_g, \sigma^2\right)\mathrm{d}\theta \\ &= \sum_j \int_{-\infty}^{+\infty} \frac{\exp\left(\theta - \delta_j\right)}{1 + \exp\left(\theta - \delta_j\right)}\Phi\left(\theta|\mu_g, \sigma^2\right)\mathrm{d}\theta \end{aligned}$$

with $\Phi(\theta|\,\mu_g, \sigma^2)$ the normal distribution with mean μ_g and variance σ^2. ($\mu_g = \mu$ if $g = A$, and $\mu_g = \mu + \gamma$ if $g = B$).

Similarly, the true value of the score variance σ_S^2 was approached by the weighted average of the expected score variance in each group.

$$\sigma_S^2 = \frac{n_A E\left(\sigma_S^2|\,g = A\right) + n_B E\left(\sigma_S^2|\,g = B\right)}{n_A + n_B}$$

The expected score variance in each group was computed as follows

$$Var(S_i|\,g) = E\left(S_i^2|\,g\right) - [E(S_i|\,g)]^2$$

$$\begin{aligned} E\left(S_i^2|\,g\right) &= E\left[\left(\sum_j x_{ij}\right)^2\right] \\ &= \sum_j E\left(x_{ij}^2\right) + 2\sum_j \sum_{j',j'>j} E\left(x_{ij}x_{ij'}\right) \\ &= \sum_j P\left(X_{ij} = 1\right) + 2\sum_j \sum_{j',j'>j} P\left(X_{ij} = 1 \cup X_{ij'} = 1\right) \\ &= \sum_j \int_{-\infty}^{+\infty} P\left(X_{ij} = 1\right)\Phi\left(\theta|\mu_g, \sigma^2\right)\mathrm{d}\theta \\ &\quad + 2\sum_j \sum_{j',j>j} \int_{-\infty}^{+\infty} P\left(X_{ij} = 1\right)\Phi\left(\theta|\mu_g, \sigma^2\right)\mathrm{d}\theta_i \int_{-\infty}^{+\infty} P\left(X_{ij'} = 1\right)\Phi\left(\theta|\mu_g, \sigma^2\right)\mathrm{d}\theta \end{aligned}$$

with $\Phi(\theta|\,\mu_g, \sigma^2)$ the normal distribution with mean μ_g and variance σ^2.

Finally, all these integrals could be estimated using Gauss-Hermite quadratures.