STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4889–4904
Published online 18 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2959

Sequential analysis of latent variables using mixed-effect latent variable models: Impact of non-informative and informative missing data

Véronique Sébille[1,*,†], Jean-Benoit Hardouin[1] and Mounir Mesbah[2]

[1] *Laboratoire de Biostatistique, Faculté de Pharmacie, Université de Nantes, 1 rue Gaston Veil,*
44035 Nantes Cedex 1, France
[2] *Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, Paris VI,*
175 rue du Chevaleret, 75013 Paris, France

SUMMARY

Sequential methods allowing for early stopping of clinical trials are widely used in various therapeutic areas. These methods allow for the analysis of different types of endpoints (quantitative, qualitative, time to event) and often provide, in average, substantial reductions in sample size as compared with single-stage designs while maintaining pre-specified type I and II errors. Sequential methods are also used when analysing particular endpoints that cannot be directly measured, such as depression, quality of life, or cognitive functioning, which are often measured through questionnaires. These types of endpoints are usually referred to as latent variables and should be analysed with latent variable models. In addition, in most clinical trials studying such latent variables, incomplete data are not uncommon and the missing data process might also be non-ignorable. We investigated the impact of informative or non-informative missing data on the statistical properties of the double triangular test (DTT), combined with the mixed-effects Rasch model (MRM) for dichotomous responses or the traditional method based on observed patient's scores (S) to the questionnaire. The achieved type I errors for the DTT were usually close to the target value of 0.05 for both methods, but increased slightly for the MRM when informative missing data were present. The DTT was very close to the nominal power of 0.95 when the MRM was used, but substantially underpowered with the S method (reduction of about 23 per cent), irrespective of whether informative missing data were present or not. Moreover, the DTT using the MRM allowed for reaching a conclusion (under H_0 or H_1) with fewer patients than the S method, the average sample number for the latter increasing importantly when the proportion of missing data increased. Incorporating MRM in sequential analysis of latent variables might provide a more powerful method than the traditional S method, even in the presence of non-informative or informative missing data. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: latent variables; mixed models; double triangular test; clinical trials, missing data; quality of life

[*] Correspondence to: Véronique Sébille, Laboratoire de Biostatistique, Faculté de Pharmacie, Université de Nantes, 1 rue Gaston Veil, 44035 Nantes Cedex 1, France.
[†] E-mail: veronique.sebille@univ-nantes.fr

Received 18 April 2007
Accepted 4 May 2007

Copyright © 2007 John Wiley & Sons, Ltd.

INTRODUCTION

Sequential methods allowing for early stopping of clinical trials in case of beneficial, harmful, or no treatment effect [1, 2] are widely used in various therapeutic areas. These methods allow for the analysis of different types of endpoints (quantitative, qualitative, and time to event) and often provide, in average, substantial reductions in sample size as compared with single-stage designs (SSDs) while maintaining pre-specified type I and II errors. All these sequential methods are usually performed on endpoints that can be directly observed and quantified, such as CD4 cell counts, success rates, or overall survival. They are also used as such when analysing particular endpoints that cannot be directly observed, such as depression, quality of life (QoL), fatigue, or cognitive functioning, for instance. In practice, such endpoints are usually evaluated using self-assessment questionnaires which consist of a set of questions often called items, whose responses are frequently combined to give scores. The common practice is to work on these scores which are expected to accurately represent the endpoint being measured and are generally assumed to be normally distributed, which is not always the case. As a matter of fact, these types of endpoints are usually referred to as being latent variables and should more probably be modelled and analysed with the so-called latent variable models [3, 4]. Indeed, latent variable models are specifically aimed at the analysis of latent variables that can be considered as random variables whose realizations are not observed in contrast with other variables [4]. Such models include Item Response Theory (IRT) models, often formulated as generalized linear mixed models [5, 6], which enable to model relationships between observed and latent variables. Some of the commonly used IRT models are the Rasch model or the Birnbaum model for dichotomous responses [3, 7], and the Rating Scale model or the Partial Credit model for polytomous responses [8, 9].

To our knowledge, specific sequential methods allowing for the analysis of latent variables have not yet been much developed despite the growing use of self-reported questionnaires in clinical trials aimed at measuring and evaluating many different latent variables such as QoL in cancer trials, dementia in Alzheimer's trials, etc. The benefit of combining sequential analysis and IRT modelling using mixed Rasch models for binary items has already been studied in the context of non-comparative clinical trials and seems very promising [10]. The statistical properties of two well-known sequential methods, namely the Sequential Probability Ratio Test and the Triangular Test [11–13], were studied and compared using IRT models and traditional scores methods in simulation studies. Incorporating IRT models in sequential analysis of latent variables seemed to be a more powerful method than the method based on observed scores, and also seemed to allow for early stopping with fewer patients. However, these simulation studies only investigated non-comparative clinical trials and did not incorporate any missing data, which is unfortunately unrealistic in most clinical trials where incomplete data are not uncommon. Moreover, the process causing the omission of data might not be at random [14] and might also have some influence on the statistical properties of the sequential procedure being used in combination with IRT models.

In this paper, we investigated the impact of informative or non-informative missing data on the statistical properties of a group sequential method often used in comparative clinical trials, the double triangular test (DTT), combined with a mixed-effects IRT model, the mixed Rasch model for dichotomous responses. The statistical properties of the DTT either combined with the mixed-effects IRT model or using the observed scores were assessed and compared by simulations regarding the type I error, power, and average sample number (ASN).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

THE RASCH MODEL

The mixed Rasch model

The basic assumption for IRT models and in particular for the Rasch model is the unidimensionality property stating that the responses to the items of a questionnaire are influenced by the underlying concept we are trying to measure (e.g. QoL, fatigue, etc.), often called latent trait and denoted by θ, and often considered as a random variable.

Consider that N patients have answered a questionnaire containing J dichotomous items. Let X_{ij} be the random variable representing the response of patient i to item j with realization x_{ij}, and θ_i be the realization of the latent trait for this patient. The mixed Rasch model postulates three assumptions [3]:

(1) Given θ_i, the response variables $X_{i1}, X_{i2}, \ldots, X_{iJ}$, are mutually independent (local independence).
(2) The probability p_{ij} of response of patient i to item j, also called the item response function of the jth item, is given by

$$p_{ij} = P(X_{ij} = x_{ij}/\theta_i; \delta_j) = f(x_{ij}/\theta_i; \delta_j) = \frac{\exp\{(\theta_i - \delta_j)x_{ij}\}}{1 + \exp(\theta_i - \delta_j)}$$

where δ_j is the parameter associated with item j and is often called the difficulty parameter for item j $(j = 1, \ldots, J)$.
(3) The variables $\theta_1, \theta_2, \ldots, \theta_N$ are mutually independent with a common underlying distribution G which is often assumed to be Gaussian.

In contrast with other IRT models, in the Rasch model, a patient's total score, $S_i = \sum_{j=1}^{J} X_{ij}$, is a sufficient statistic for the latent trait θ_i.

Identifiability constraints and estimation of the parameters

Identifiability of the model can be ensured by putting one constraint on the parameters. Usually, we assume that the mean of the latent trait is 0 or that the sum of the item parameters $\sum_j \delta_j = 0$.

Let $\delta = (\delta_1, \delta_2, \ldots, \delta_J)$. The marginal likelihood of the mixed Rasch model is given by

$$L(\delta, \mu, \sigma^2/x) = \prod_{i=1}^{N} \int_{\Re} \prod_{j=1}^{J} P(X_{ij} = x_{ij}/\theta_i; \delta_j) \cdot G(\theta_i/\mu, \sigma^2)\, \mathrm{d}\theta_i$$

where $G(./\mu, \sigma^2)$ is the Gaussian distribution function with mean μ and variance σ^2.

The person parameters (latent traits) can be jointly estimated with the item parameters by marginal maximum likelihood (MML) estimation obtained from integrating out the random effects. The MML estimators that are obtained are asymptotically efficient [3, 15, 16].

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

SEQUENTIAL ANALYSIS

Observed scores

Consider a comparative clinical trial that involves the comparison of responses to a self-assessment questionnaire aimed at measuring QoL or fatigue, for instance, in two parallel treatment groups (group 1 and group 2). Let us assume that we are working on summation scores obtained from the responses and that they follow some distribution assumed to be Gaussian with unknown parameters μ_1 (in group 1) and μ_2 (in group 2) and common σ^2. Let the difference between treatment groups be parameterized as $\varphi = (\mu_2 - \mu_1)/\sigma$.

The test statistics Z and V

We are testing the null hypothesis H_0: $\varphi = 0$ against the two-sided alternative H_1: $\varphi \neq 0$ with H_1^+: $\varphi > 0$ and H_1^-: $\varphi < 0$. The log-likelihood can be expressed according to both independent samples, and its derivatives can be used to derive the test statistics Z and V, both evaluated under the null hypothesis. The test statistic Z is the efficient score for φ depending on the observed scores, and the test statistic V is Fisher's information for φ. Their expressions for a normally distributed endpoint are given in Appendix and details of the computations are described at length by Whitehead [1].

Latent variables

We shall now consider the latent case, i.e. the case where the latent trait θ_{ig} of each patient i is unobserved in each treatment group g $(g = 1, 2)$. Let us assume that $n_1 + n_2$ data have been gathered so far in the two treatment groups and that the data form two separate sequences of independent observations: $(\theta_{11}, \theta_{21}, \ldots, \theta_{n_1 1}) \sim f_{\theta_1}(\psi_1, \eta)$ and $(\theta_{12}, \theta_{22}, \ldots, \theta_{n_2 2}) \sim f_{\theta_2}(\psi_2, \eta)$, where ψ_1 and $\psi_2 \in \Re$ and where η is an unknown common nuisance parameter (possibly vector valued). Following Whitehead's notations [1], let $\varphi = (\psi_2 - \psi_1)/2$ be the parameter of interest and the nuisance parameter be made up of $\phi = (\psi_1 + \psi_2)/2$ and η. Hence, $\psi_1 = \phi - \varphi$ and $\psi_2 = \phi + \varphi$ and $\theta_1 \sim f_{\theta_1}(\phi - \varphi, \eta)$ and $\theta_2 \sim f_{\theta_2}(\phi + \varphi, \eta)$. The hypotheses we are testing can be expressed as H_0: $\psi_1 = \psi_2$ against H_1: $\psi_1 \neq \psi_2$ or as H_0: $\varphi = 0$ against H_1: $\varphi \neq 0$. The log-likelihood of φ, ϕ, and η can be written as

$$l(\varphi, \phi, \eta) = l^{(1)}(\psi_1, \eta) + l^{(2)}(\psi_2, \eta)$$

Assuming a Rasch model for patients' item responses, and $\eta = (\sigma^2, \delta_1, \ldots, \delta_J)$, where σ^2 is the common variance of the latent traits θ_1 and θ_2 in each treatment group and $\delta_1, \ldots, \delta_J$ are the J item parameters, we can write

$$l^{(g)}(\psi_g, \sigma^2, \delta_1, \ldots, \delta_J) = \sum_{i=1}^{n_g} \log \left\{ \int f_{\theta_g}(\psi_g, \sigma^2) \prod_j \frac{\mathrm{e}^{(\theta_{ig} - \delta_j) x_{ijg}}}{1 + \mathrm{e}^{(\theta_{ig} - \delta_j)}} \, \mathrm{d}\theta_{ig} \right\}, \quad g = 1, 2$$

The test statistics Z and V can be derived in the following way under H_0:

$$Z = \frac{\partial l}{\partial \varphi}(\varphi, \phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)\bigg|_{\varphi=0} \quad \text{and} \quad V = -\frac{\partial^2 l}{\partial \varphi^2}(\varphi, \phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)\bigg|_{\varphi=0}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

where ϕ^* and $\eta^* = (\sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)$ are the maximum likelihood estimates of ϕ and η under the assumption that both series of data come from the same distribution.

Since $l(\varphi, \phi, \eta) = l^{(1)}(\psi_1, \eta) + l^{(2)}(\psi_2, \eta)$ and $\psi_1 = \phi - \varphi$ and $\psi_2 = \phi + \varphi$, we can write

$$\begin{aligned}\frac{\partial l}{\partial \varphi}(\varphi, \phi, \eta) &= \frac{\partial l^{(1)}}{\partial \varphi}(\psi_1, \eta) + \frac{\partial l^{(2)}}{\partial \varphi}(\psi_2, \eta)\\ &= \frac{\partial \psi_1}{\partial \varphi}\cdot\frac{\partial l^{(1)}}{\partial \psi_1}(\psi_1, \eta) + \frac{\partial \psi_2}{\partial \varphi}\cdot\frac{\partial l^{(2)}}{\partial \psi_2}(\psi_2, \eta)\\ &= \frac{\partial l^{(2)}}{\partial \psi_2}(\psi_2, \eta) - \frac{\partial l^{(1)}}{\partial \psi_1}(\psi_1, \eta)\end{aligned}$$

Hence, under $H_0(\varphi = 0)$:

$$Z = \frac{\partial l^{(2)}}{\partial \psi_2}(\phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*) - \frac{\partial l^{(1)}}{\partial \psi_1}(\phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)$$

The test statistic V can be approximated by the following expression [1] for planning purposes when the two samples are large, of about the same size and when φ is small:

$$\begin{aligned}V &= -\frac{\partial^2 l}{\partial \varphi^2}(\varphi, \phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)\Big|_{\varphi=0}\\ &= -\frac{\partial^2 l^{(2)}}{\partial \psi_2^2}(\phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*) - \frac{\partial^2 l^{(1)}}{\partial \psi_1^2}(\phi^*, \sigma^{2*}, \delta_1^*, \ldots, \delta_J^*)\end{aligned}$$

Estimation of the test statistics Z and V is done by maximizing the marginal likelihood obtained from integrating out the random effects. Quasi-Newton procedures can be used to maximize the likelihood, along with adaptive Gaussian quadrature to integrate out the random effects [17].

The double triangular test

The DTT (Figure 1) uses a sequential plan defined by two perpendicular axes [1]. The horizontal axis corresponds to the test statistic V, which represents the quantity of information accumulated since the beginning of the trial (Fisher's information for the parameter of interest). The vertical axis corresponds to the test statistic Z, which represents the benefit as compared with the null hypothesis (efficient score for the parameter of interest). For the DTT, two single triangular tests, symmetrical about the V axis, are combined: the continuation region is situated inside of the two triangles, the region of non-rejection of the null hypothesis is situated between the two inner boundaries, the region of rejection of the null hypothesis in favour of H_1^- is situated beneath the lower outer boundary, and the region of rejection of the null hypothesis in favour of H_1^+ is situated above the upper outer boundary. The boundaries depend on the statistical hypotheses (values of the expected treatment benefit, α and β) and on the number of subjects included between two analyses. They can be adapted at each analysis when this number varies from one analysis to the other, using the 'Christmas tree' correction [1]. At each analysis, Z and V are computed from all

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

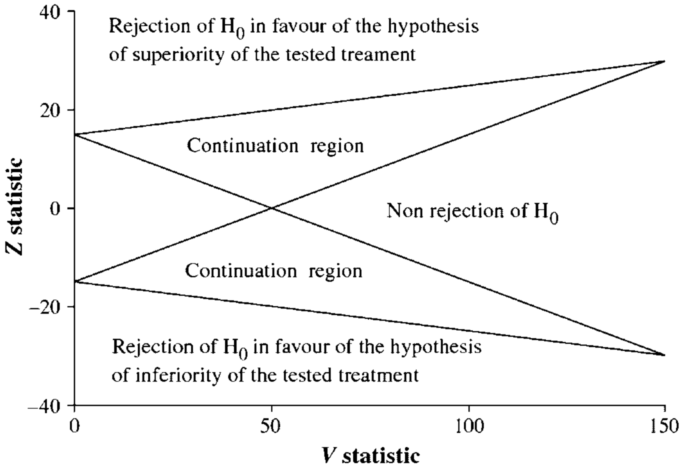

Figure 1. Stopping boundaries based on the double triangular test (DTT) for $\alpha = 0.05$ and $\beta = 0.025$ with an effect size (reference improvement) $= 0.5$.

the available data and Z is plotted against V, thus defining a sample path. The trial is continued as long as the sample path remains in the continuation region. A conclusion is reached as soon as the sample path crosses one of the boundaries of the test. In practice, a computer software called PEST [18] is available and can be used for the planning, monitoring, and analysis of sequential comparative clinical trials.

Simulations

The latent trait θ was considered as a random variable following a normal distribution: $\theta \sim \mathrm{N}(0, 1)$, and θ_i represents the latent trait of the ith patient. For each patient, the probability of responding to each item was computed according to the Rasch model:

$$p_{ij} = P(X_{ij} = x_{ij}/\theta_i; \delta_j) = \frac{\exp\{(\theta_i - \delta_j)x_{ij}\}}{1 + \exp(\theta_i - \delta_j)}$$

where $x_{ij} = 0$ for a negative response and $x_{ij} = 1$ for a positive response. In order to assess the impact of model misspecification, the probability of responding to each item was also computed according to the Birnbaum model, which is a generalization of the Rasch model:

$$p_{ij} = P(X_{ij} = x_{ij}/\theta_i; \delta_j, a_j) = \frac{\exp\{a_j(\theta_i - \delta_j)x_{ij}\}}{1 + \exp\{a_j(\theta_i - \delta_j)\}}$$

where a_j is called the discriminating power.

Three missing data mechanisms have been described by Rubin [14]: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). For instance, in case of a self-reported QoL questionnaire, data can be considered MCAR if the probability of

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

having a missing data (missing response on one or more items, for instance) is independent of the patient's QoL. Data will be considered MAR if the probability of missing data may depend on the previous patient's QoL but not on its present or future (unobserved) QoL. In contrast, data will be considered MNAR if the probability of missing data depends on the patient's present and future (unobserved) QoL. Methods for identifying the different types of missingness and hence determining appropriate methods of analysis have been discussed elsewhere in detail, especially in the longitudinal setting [19–22].

Data were simulated mainly according to two different mechanisms: MCAR and MNAR. Another latent variable denoted by ξ was used, corresponding to non-response propensity, which represents the tendency of not responding, varying between individuals. This latent variable may be influenced by the patient's latent trait θ (QoL, fatigue, pain, anxiety, etc.) and may thus involve an informative non-response framework corresponding to MNAR data. To simulate the missing values, we assumed that each patient had a non-response propensity to each item represented by the latent variable ξ that followed a normal distribution with zero mean and variance unity. The correlation coefficient between θ and ξ was denoted as ρ. Let π be the expected rate of missing values for each item and π_i be the probability for the ith patient to have a missing value to each item. In case of missing data, this probability was assumed to have a lower bound equal to 1 per cent and to be centred on π (for π between 2 and 50 per cent).

Let

$$\xi_i^* = \begin{cases} -2 & \text{if } \xi_i \leqslant -2 \\ \xi_i & \text{if } -2 < \xi_i < 2 \\ 2 & \text{if } \xi_i \geqslant 2 \end{cases} \quad \text{and} \quad \pi_i = \frac{\xi_i^*(\pi - 0.01)}{2} + \pi$$

According to the value of ρ, the missing values will be considered as being non-informative (MCAR) or informative (MNAR): for $\rho = 0$, the data will be MCAR, for $\rho \neq 0$, the data will not be considered as MCAR anymore but will be considered MNAR. We assumed that a patient with a low level on the latent trait (low level of QoL, for instance) had a higher propensity of absence of response to the items, so ρ was assumed to be $\leqslant 0$.

A thousand comparative clinical trials were simulated. The latent trait in the control group θ_1 was assumed to follow a normal distribution with mean $\mu_1 = 0$ and variance $\sigma^2 = 1$, and the latent trait in the experimental group θ_2 was assumed to follow a normal distribution with mean $\mu_2 = \mu_1 + d$ and same variance. The trial involved the comparison of the two hypotheses: H_0: $d = 0$ against H_1: $d \neq 0$. We assumed that a five-item scale (that could represent one of the dimensions of a questionnaire, like for, instance, physical, social, or emotional dimension) was used. For the Rasch and the Birnbaum models, the corresponding item parameters were assumed to be forming a part of a calibrated item bank [23]: $\delta_1 = -1.0$, $\delta = -0.5$, $\delta_3 = 0.0$, $\delta_4 = 0.5$, and $\delta_5 = 1.0$. For the Birnbaum model, the parameters a_j were randomly drawn in the interval [0.5–2.0] with a median of 1. Data were simulated with four different values for ρ: $\rho = 0$ (MCAR data), -0.4, -0.7, and -0.9 (MNAR data). The probability of non-response for each item is represented in Figures 2(a)–(d) as a function of the latent trait θ, for different values of ρ (each dot represents an individual).

We compared the use of mixed Rasch modelling methods with summation scores methods using the DTT. The sequential analyses were performed for every 40 included patients, the effect size was equal to $d = 0.5$ under H_1, and $\alpha = \beta = 0.05$ for all simulations.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

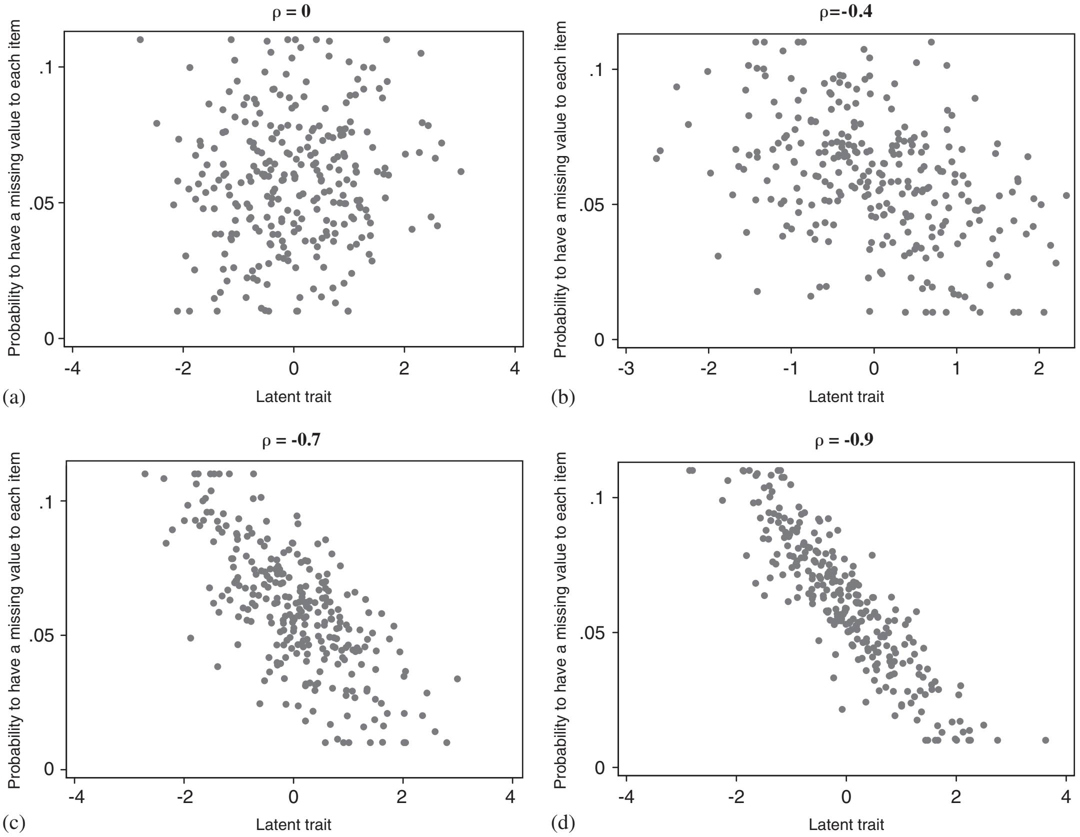

Figure 2. Probability of having a missing value to each item as a function of the latent trait θ for different values of the correlation coefficient ρ between θ and the non-response propensity ξ for $\pi = 5$ per cent: (Panel a) $\rho = 0$; (Panel b) $\rho = -0.4$; (Panel c) $\rho = -0.7$; and (Panel d) $\rho = -0.9$.

SIMULATION RESULTS

Table I (data simulated under a Rasch model) shows the type I error and the power for the DTT for different proportions π of missing data and for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ, using the method based on observed scores or the mixed Rasch model.

The type I errors were usually close to the target value of 0.05 for both methods when $\pi \leqslant 10$ per cent, but slightly increased when $\pi = 15$ or 20 per cent for the mixed Rasch model. Informative data ($\rho \neq 0$) did not seem to affect the significance level of the method based on observed scores, whereas it was somewhat increased for the mixed Rasch model (from 0.053 when $\rho = 0$ to 0.061 when $\rho = -0.9$, on average). The DTT was very close to the nominal power of 0.95 when the mixed Rasch model was used, but substantially underpowered when the method based on observed scores was used. Indeed, for the method based on observed scores, as compared with the target power value of 0.95, there were decreases in power of approximately 23 per cent for all values of

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table I. Type I error and power for the double triangular test (DTT) using the method based on observed scores or the mixed Rasch model for various proportions π of missing data and for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ (nominal $\alpha=\beta=0.05$, five-item scale, 1000 simulations, data simulated under a Rasch model).

		Type I error/Power	
π	ρ	Scores	Rasch model
0	0	0.040/0.712	0.051/0.952
0.02	0	0.053/0.708	0.057/0.946
	−0.4	0.064/0.708	0.051/0.946
	−0.7	0.055/0.714	0.069/0.949
	−0.9	0.049/0.703	0.056/0.936
0.04	0	0.052/0.734	0.046/0.943
	−0.4	0.049/0.711	0.038/0.960
	−0.7	0.055/0.719	0.065/0.952
	−0.9	0.042/0.757	0.057/0.966
0.06	0	0.050/0.732	0.042/0.943
	−0.4	0.039/0.754	0.045/0.948
	−0.7	0.049/0.736	0.048/0.962
	−0.9	0.054/0.708	0.061/0.949
0.08	0	0.042/0.749	0.048/0.947
	−0.4	0.041/0.765	0.062/0.949
	−0.7	0.040/0.756	0.050/0.947
	−0.9	0.052/0.755	0.058/0.963
0.10	0	0.046/0.736	0.046/0.949
	−0.4	0.043/0.744	0.055/0.945
	−0.7	0.056/0.708	0.063/0.968
	−0.9	0.045/0.728	0.064/0.976
0.15	0	0.046/0.769	0.068/0.932
	−0.4	0.039/0.754	0.052/0.955
	−0.7	0.043/0.725	0.057/0.956
	−0.9	0.028/0.713	0.058/0.972
0.20	0	0.049/0.796	0.068/0.946
	−0.4	0.045/0.761	0.062/0.960
	−0.7	0.045/0.730	0.063/0.981
	−0.9	0.051/0.695	0.072/0.975

π and ρ considered. More precisely, concerning the π and ρ effects, the power seemed to increase slightly with π, the increase depending on the level of ρ for both methods, but in a different way: the increase was observed for the method based on observed scores when $\rho=0$ or -0.4 (the power increased on average, from 0.712 when $\pi=0$ per cent to 0.779 when $\pi=20$ per cent for the method based on observed scores, and from 0.952 when $\pi=0$ per cent to 0.953 when $\pi=20$ per cent for the mixed Rasch model), whereas it was observed for the mixed Rasch model when

Copyright © 2007 John Wiley & Sons, Ltd.

Table II. Sample size for the single-stage design (SSD) and average sample number (ASN) required to reach a conclusion under H_0 and H_1 for the double triangular test (DTT) using the method based on observed scores or the mixed Rasch model for various proportions π of missing data and for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ (nominal $\alpha=\beta=0.05$, effect size$=0.5$, five-item scale, 1000 simulations, data simulated under a Rasch model).

			DTT*	Scores	Rasch model
π	ρ	SSD	H_0/H_1	H_0/H_1	H_0/H_1
0	0	208	149/125	149/154	148/125
0.02	0	208	149/125	169/171	154/129
	−0.4	208	149/125	168/173	157/132
	−0.7	208	149/125	169/175	155/130
	−0.9	208	149/125	170/172	155/130
0.04	0	208	149/125	184/183	156/131
	−0.4	208	149/125	183/185	156/130
	−0.7	208	149/125	183/181	159/128
	−0.9	208	149/125	185/184	159/127
0.06	0	208	149/125	199/191	158/132
	−0.4	208	149/125	198/190	157/130
	−0.7	208	149/125	196/192	161/127
	−0.9	208	149/125	198/191	159/131
0.08	0	208	149/125	213/198	160/135
	−0.4	208	149/125	210/197	158/130
	−0.7	208	149/125	210/200	158/129
	−0.9	208	149/125	212/196	159/127
0.10	0	208	149/125	232/225	161/136
	−0.4	208	149/125	231/220	161/129
	−0.7	208	149/125	233/222	164/130
	−0.9	208	149/125	233/222	161/125
0.15	0	208	149/125	274/253	167/139
	−0.4	208	149/125	275/247	169/135
	−0.7	208	149/125	273/248	166/131
	−0.9	208	149/125	275/248	165/127
0.20	0	208	149/125	332/287	172/144
	−0.4	208	149/125	333/286	174/137
	−0.7	208	149/125	333/285	174/127
	−0.9	208	149/125	330/286	172/128

*ASN for the DTT for a normally distributed endpoint provided by PEST software.

$\rho=-0.7$ or -0.9 (on average, from 0.712 when $\pi=0$ per cent to 0.713 when $\pi=20$ per cent for the method based on observed scores and from 0.952 when $\pi=0$ per cent to 0.978 when $\pi=20$ per cent for the mixed Rasch model).

Table II (data simulated under a Rasch model) shows the ASN of the number of patients required to reach a conclusion under H_0 and H_1 for the DTT for different proportions π of missing data and

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ using the method based on observed scores or the mixed Rasch model. We also computed for comparison purposes the number of patients required by a two-sided SSD and the approximate ASN for the DTT computed with PEST software [18] when a normally distributed endpoint is assumed when planning the trial. As expected, the ASNs were smaller for the DTT as compared with the sample size required by the SSD using either method (observed scores or mixed Rasch model) when no missing data were present ($\pi = 0$), and similar to the ASN computed using PEST software for the DTT for a normally distributed endpoint under H_0. The same feature was also observed under H_1, except for the method based on observed scores, which displayed a higher ASN than the others (154 instead of 125). Moreover, the ASNs increased as the proportion π of missing data increased for almost all values of ρ, particularly when using the method based on observed scores. Indeed, for this method using the DTT, as compared with the number of patients required by the SSD, as π increased from 0 to 20 per cent, the ASN under H_0 (H_1) ranged from a decrease in the number of patients required to reach a conclusion of -28 per cent (-26 per cent) to an increase in this number of $+60$ per cent ($+38$ per cent), for all values of ρ. In contrast, for the method based on the mixed Rasch model, as compared with the SSD, when π increased from 0 to 20 per cent of missing data, the decreases in sample size obtained with the DTT under H_0 diminished only slightly, ranging from -29 to -17 per cent. Under H_1, the effect of π was more

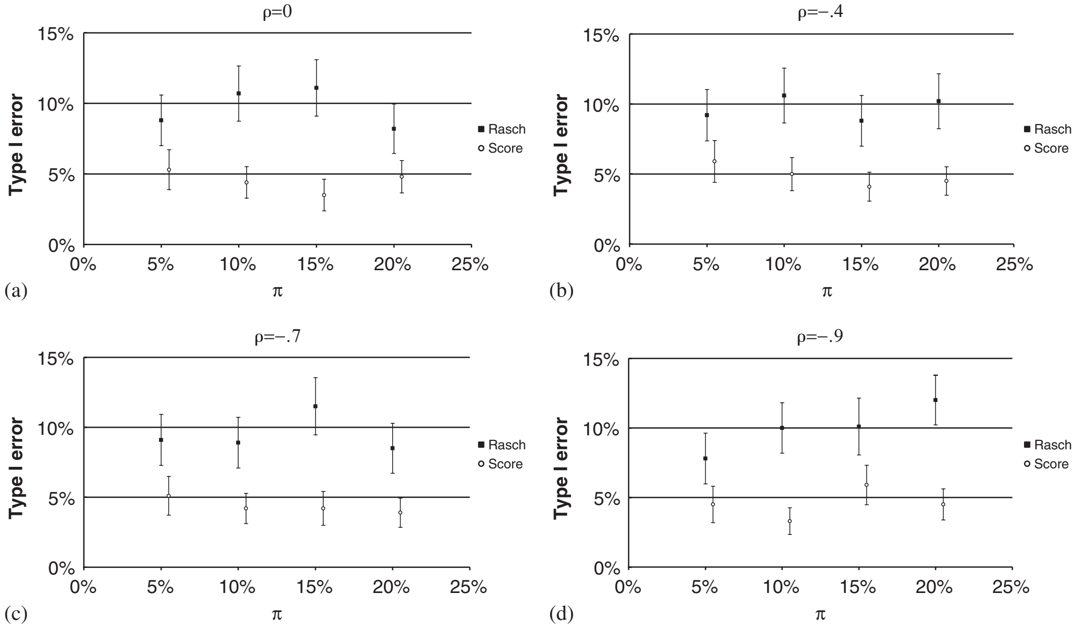

Figure 3. Type I error probability achieved by the single-stage design (SSD) using the observed scores (empty circles) or the mixed Rasch model (full squares) as a function of the proportion π of missing data for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ (data simulated under the Birnbaum model): (Panel a) $\rho = 0$; (Panel b) $\rho = -0.4$; (Panel c) $\rho = -0.7$; and (Panel d) $\rho = -0.9$. The 95 per cent confidence intervals were calculated using the normal approximation to the binomial distribution.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

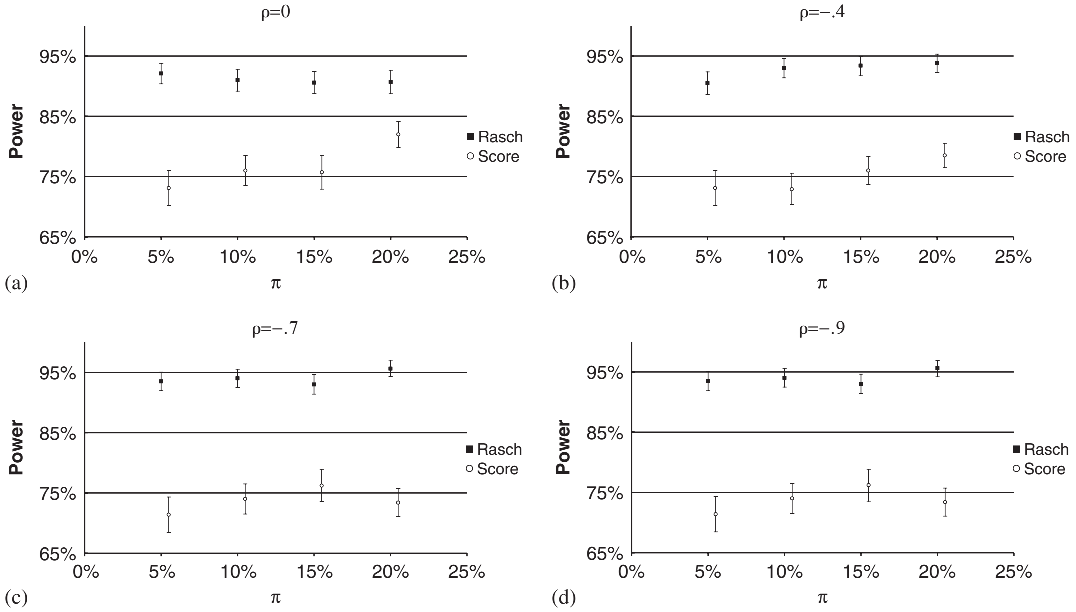

Figure 4. Power achieved by the single-stage design (SSD) using the observed scores (empty circles) or the mixed Rasch model (full squares) as a function of the proportion π of missing data for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ (data simulated under the Birnbaum model): (Panel a) $\rho=0$; (Panel b) $\rho=-0.4$; (Panel c) $\rho=-0.7$; and (Panel d) $\rho=-0.9$. The 95 per cent confidence intervals were calculated using the normal approximation to the binomial distribution.

marked when $\rho=0$ or $\rho=-0.4$, where the decreases in sample size obtained with the DTT varied from −40 to −32 per cent as π increased from 0 to 20 per cent, whereas it remained stable when $\rho=-0.7$ or $\rho=-0.9$ at about −38 per cent for all values of π.

Correspondingly, as compared with the ASN calculated for the DTT for a normally distributed endpoint using PEST software, for the method based on observed scores, as π increased from 0 to 20 per cent, the increase in ASN under H_0 (H_1) ranged from no increase (+23 per cent) to +123 per cent (+129 per cent) for all values of ρ. In contrast, for the method based on the mixed Rasch model, as compared with the ASN calculated for the DTT for a normally distributed endpoint using PEST software, the increase in ASN under H_0 (H_1) ranged from no increase (no increase) to +38 per cent (+13 per cent for $\rho=0$ or $\rho=-0.4$ and +2 per cent for $\rho=-0.7$ or $\rho=-0.9$), as π increased from 0 to 20 per cent.

Figures 3 and 4 (data simulated under a Birnbaum model) show type I error probability and power, respectively, achieved by the SSD using the observed scores or the mixed Rasch model as a function of the proportion π of missing data for different values of the correlation coefficient ρ between the latent trait θ and the non-response propensity ξ. The significance level achieved by the DTT seemed to be unaffected with the method based on observed scores, whereas it was increased when the mixed Rasch model was used for all values of π and ρ (the mean type I error was 0.098 with the latter and 0.047 with the former), the increase being more marked when $\rho=-0.9$. The DTT was a bit underpowered when the mixed Rasch model was used (0.929 on average), and

Copyright © 2007 John Wiley & Sons, Ltd.

still substantially underpowered when the method based on observed scores was used (0.741 on average), the power slightly increasing for both methods with π (especially when $\rho = 0$ or -0.4 for the method based on observed scores) and with ρ for the mixed Rasch model. The ASNs displayed the same features as the ones already observed when the data were simulated under the Rasch model (data not shown): an important increase was observed as π increased, mostly for the method based on observed scores, the increase being more moderate for the mixed Rasch model.

DISCUSSION

We investigated the impact of informative or non-informative missing data on the statistical properties of the DTT, combined with a mixed-effects IRT model, the mixed Rasch model or the traditional approach based on observed scores.

Simulation studies showed that for the DTT: (i) the type I error α was correctly maintained for both methods when non-informative data were present ($\rho = 0$) for almost all values of the proportion π of missing data, whereas it was increased with $|\rho|$ for the mixed Rasch model; (ii) the power of the DTT was accurately maintained for the mixed Rasch model, but it was substantially underpowered with the method based on observed scores for all values of π and ρ even if a slight increase was observed for both methods as π increased; and (iii) as expected using group sequential analysis, both methods allowed substantial reductions in ASNs as compared with the SSD, the largest reduction being observed with the mixed Rasch model; the reduction in sample size diminished importantly mostly for the method based on observed scores as π increased and the benefit of using the DTT was completely lost for this method, the ASN being similar or more often larger than the sample size required by the SSD under both H_0 and H_1. Finally, model mis-specification for the Rasch model (when data were simulated with a Birnbaum model) had an impact mostly on the type I error α, but it was more moderate on the power. Thus, this illustrates that mixed Rasch models should be used only when they provide a good fit to the data and that other models should be investigated otherwise.

The inflation of the type I error α for the mixed Rasch model in the presence of informative data ($\rho \neq 0$) and the important loss in power of the DTT based on the observed scores method as compared with the mixed Rasch model might be explained by looking at the distributions of the test statistics Z and V computed using both methods. According to asymptotic distributional results, we might expect the sequences of test statistics $(Z_1, Z_2, \ldots, Z_K)$ to be multivariate normal, with $Z_k \sim \mathrm{N}(\mathrm{ES} * V_k, V_k)$, where ES denotes the effect size (equal to 0.5 in the simulations under H_1), for $k = 1, 2, \ldots, K$ analyses [1, 2]. We looked at the distribution of the standardized test statistics

$$Z'_4 = \frac{Z_4 - (\mathrm{ES} * V_4)}{\sqrt{V_4}} \sim \mathrm{N}(0, 1)$$

(corresponding to the fourth sequential analysis performed on 160 patients) under H_0 and H_1 computed using the method based on observed scores or the mixed Rasch models for $\pi = 0$ and $\pi = 8$ per cent for $\rho = 0$, -0.4, and -0.9. The normality assumption for Z'_4 was not rejected using a Kolmogorov–Smirnov test, whatever the method used to compute the test statistics, for all values of π and ρ. However, the test statistics Z'_4 computed using the mixed Rasch model increased with ρ in absolute value, its normal distribution being centred above zero when $|\rho| > 0$ (at 0.12 for $\rho = -0.9$, for instance), thus explaining to a certain extent the inflation of the type I error α in

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

the presence of informative data for this model. Furthermore, concerning the power of the DTT, the standardized test statistic Z_4' computed using the method based on observed score was always significantly lower than the standardized test statistic Z_4' computed using the method based on the mixed Rasch model for all values of π and ρ considered, its sample mean being significantly lower than zero under all circumstances. In addition, the moderate increase in power observed as π increased, especially with the method based on observed scores, can be directly related to the inflation of the corresponding ASNs and number of sequential analyses to be performed (from about four sequential analyses when $\pi = 0$ per cent to more than eight when $\pi = 20$ per cent).

The missing data mechanism simulated with the use of the correlation coefficient ρ between the latent trait θ and the propensity of non-response ξ did seem to alter the statistical properties of the testing procedures, especially in terms of significance level for the mixed Rasch model. It is well known that performing analyses on informative data without incorporating the missing data process can also lead to bias in treatment effect estimates [19]. Several methods have been proposed to model the missing data mechanism [20, 24], including, more recently, a method based on IRT modelling [25], where possible bias in the item parameter estimates was investigated by simulating about the same informative data mechanisms as ours. The bias could sometimes be important and could be reduced by taking into account the missing data mechanism modelled with a mixed-effects IRT model. This type of strategy could also be incorporated in a group sequential analysis setting, and more work is needed.

Finally, we worked only on binary items, whereas polytomous items appear more frequently in most questionnaires used in clinical trial practice. Other mixed-effects IRT models such as the Partial Credit model or the Rating Scale model [8, 9] might be more appropriate in this context and are currently being investigated.

It has been reported that latent variable models might provide more accurate assessment of health status as compared with observed scores [26, 27]. Hence, if an IRT model, such as the mixed Rasch model for dichotomous responses, shows a good fit to the data, incorporating mixed-effects IRT models in sequential analysis of latent variables will provide a more powerful method to detect therapeutic effects than the traditional method based on observed scores, even in the presence of non-informative or informative missing data.

APPENDIX

The test statistics Z and V for a normally distributed endpoint are given by

$$Z = \frac{n_1 n_2}{(n_1+n_2)\cdot D}(\bar{s}_2 - \bar{s}_1) \quad \text{and} \quad V = \frac{n_1 n_2}{(n_1+n_2)} - \frac{Z^2}{2(n_1+n_2)}$$

in which:

- n_g is the cumulated number of patients (since the beginning of the trial) in group g ($g = 1, 2$),
- $\bar{s}_g = \sum_{j=1}^{n_g} s_{gj}/n_g$ where s_{gj} denotes the observed scores of patient j in group g,
- D is the maximum likelihood estimate of σ under the null hypothesis:

$$D = \sqrt{\frac{Q}{n_1+n_2} - \left(\frac{R}{n_1+n_2}\right)^2}$$

Copyright © 2007 John Wiley & Sons, Ltd.

with

$$Q=\sum_{j=1}^{n_1} s_{1j}^2+\sum_{j=1}^{n_2} s_{2j}^2 \quad \text{and} \quad R=\sum_{j=1}^{n_1} s_{1j}+\sum_{j=1}^{n_2} s_{2j}$$

Details of the computations are described at length by Whitehead [1].

REFERENCES

1. Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (2nd edn). Wiley: Chichester, 1997.
2. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC: Boca Raton, FL, 1999.
3. Fisher GH, Molenaar IW. *Rasch Models*, *Foundations*, *Recent Developments*, *and Applications*. Springer: New York, 1995.
4. Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling*. Chapman & Hall/CRC: Boca Raton, FL, 2004.
5. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
6. McCulloch CE, Searle SR. *Generalized*, *Linear*, *and Mixed Models*. Wiley: New York, 2000.
7. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests* (expanded edn). Nielsen & Lydiche/The University of Chicago Press: Copenhagen, Chicago, 1980.
8. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**:561–573.
9. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; **47**:149–174.
10. Sébille V, Mesbah M. Sequential analysis of quality of life Rasch measurements. In *Probability*, *Statistics and Modelling in Public Health*, Nikulin M, Commenges D, Huber C (eds). Springer: Berlin, 2006; 421–439.
11. Wald A. *Sequential Analysis*. Wiley: New York, 1947.
12. Whitehead J, Jones DR. The analysis of sequential clinical trials. *Biometrika* 1979; **66**:443–452.
13. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983; **39**:227–236.
14. Rubin DB. Inference, missing data. *Biometrika* 1976; **63**:581–592.
15. Thissen D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* 1982; **47**:175–186.
16. Hamon A, Mesbah M. Questionnaire reliability under the Rasch model. In *Statistical Methods for Quality of Life Studies*: *Design*, *Measurements and Analysis*, Mesbah M, Cole BF, Lee MLT (eds). Kluwer: Amsterdam, 2002.
17. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
18. MPS Research Unit. *PEST 4*: *Operating Manual*. The University of Reading: Reading, MA, 2000.
19. Choi S, Lu IL. Effect of non-random missing data mechanisms in clinical trials. *Statistics in Medicine* 1995; **14**:2675–2684.
20. Little RJA. Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112–1121.
21. Curran D, Bacchi M, Hsu Schmitz SF, Molenberghs G, Sylvester RJ. Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine* 1998; **17**:739–756.
22. Fitzmaurice GM, Lipsitz SR, Molenberghs G, Ibrahim JG. A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *Journal of the Royal Statistical Society, Series A* 2005; **168**:723–735.
23. Holman R, Lindeboom R, Glas CAW, Vermeulen M, de Haan RJ. Constructing an item bank using item response theory: the AMC linear disability score project. *Health Services and Outcomes Research Methodology* 2003; **4**:19–33.
24. Scharfstein D, Ronitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-response models (with Discussion). *Journal of the American Statistical Association* 1999; **94**:1096–1146.
25. Holman R, Glas CAW. Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology* 2005; **58**:1–17.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

26. McHorney CA, Haley SM, Ware Jr JE. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology* 1997; **50**:451–461.
27. Kosinski M, Bjorner JB, Ware Jr JE, Batenhorst A, Cady RK. The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: a re-analysis of three clinical trials. *Quality of Life Research* 2003; **12**:903–912.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim