1

Sequential Analysis of Quality of Life Measurements Using Mixed Rasch Models

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. Email: veronique.sebille@univ-nantes.fr. Tel: +33 (0)2 40 41 29 96, Fax: +33 (0)2 40 41 28 29

2. Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, Paris VI, 175 rue du Chevaleret, 75013 Paris, France

Abstract: Early stopping of clinical trials in case of either beneficial or deleterious effect of a treatment on quality of life (QoL) is an important issue. QoL is usually evaluated using self-assessment questionnaires and responses to the items are usually combined into QoL scores assumed to be normally distributed. However, these QoL scores are rarely normally distributed and usually do not satisfy a number of basic measurement properties. An alternative is to use item response theory (IRT) models such as the Rasch model for binary items which takes into account the categorical nature of the items. In this framework, the probability of response of a patient on an item depends upon different kinds of parameters: the "ability level" of the person (which reflects his/her current QoL) and a set of parameters characterizing each items.

Sequential analysis and mixed Rasch models assuming either known or unknown items parameters values were combined in the context of phase II, phase III comparative clinical trials. The statistical properties of the Triangular Test (TT) were compared using mixed Rasch models and the traditional method based on QoL scores by means of simulations.

The type I error of the TT was correctly maintained for the methods based on Qol scores and the Rasch model assuming known items parameters values, but was higher than expected when items parameters were assumed to be unknown. The power of the TT was satisfactorily maintained when Rasch models were used but the test was underpowered when the QoL scores method was used. All methods allowed substantial reductions in average sample numbers as compared with fixed sample designs, especially the method based on Rasch models. The use of IRT models in sequential analysis of QoL endpoints seems to provide a more powerful method to detect therapeutic effects than the traditional QoL scores method and to allow for reaching a conclusion with fewer patients.

Keywords and phrases: Quality of life; Item Response Theory; Rasch models; Triangular Test; Clinical Trials; Mixed Models

---

1.1 INTRODUCTON

Many clinical trials attempt to measure Health-Related Quality of Life (QoL) which refers to "the extent to which one's usual or expected physical, emotional and social well-being are affected by a medical condition or its treatment" [Cella (1995), Fairclough (2002)]. Early stopping of clinical trials either in case of beneficial or deleterious effect of treatment on QoL is an important issue. However, each domain of health can have several components (e.g., symptoms, ability to function, disability) and translating these various domains of health into quantitative values to measure quality of life is a complex task, drawing from the field of psychometrics, biostatistics, and clinical decision theory. In clinical trials in which specific therapeutic interventions are being studied, patient's QoL is usually evaluated using self-assessment questionnaires which consist of a set of questions called items (which can be dichotomous or polytomous) which are frequently combined to give scores. The common practice is to work on average scores which are generally assumed to be normally distributed. However, these average scores are rarely normally distributed and usually do not satisfy a number of basic measurement properties including sufficiency, unidimensionality, or reliability. An alternative is to use item response theory (IRT) models [Fisher and Molenaar (1995)], such as the Rasch model for binary items, which takes into account the categorical nature of the items by introducing an underlying response model relating those items to a latent parameter interpreted as the true individual QoL.

Early stopping of a trial can occur either for efficacy, safety or futility reasons. Several early termination procedures have been developed to allow for repeated statistical analyses on accumulating data and for stopping a trial as soon as the information is sufficient to conclude. Among the sequential methods that have been developed over the last few decades [Pocock (1997), O'Brien and Fleming (1979), Lan and De Mets (1983)], the Sequential Probability Ratio Test (SPRT) and the Triangular Test (TT), which were initially developed by Wald (1947) and Anderson (1960) and later extended by Whitehead to allow for sequential analyses on groups of patients [Whitehead and Jones (1979), Whitehead and Stratton (1983)] have some of the interesting following features. They allow for: (i) early stopping under H_0 or under H_1, (ii) the analysis of

quantitative, qualitative or censored endpoints, (iii) type I and II errors to be correctly maintained at their desired planning phase values, (iv) substantial sample size reductions as compared with the single-stage design (SSD).

While sequential methodology is often used in clinical trials, IRT modelling, as a tool for scientific measurement, is not quite well established in the clinical trial framework despite a number of advantages offered by IRT to analyze clinical trial data [Holman, Glas and Haan (2003)]. Moreover, it has been suggested that IRT modelling offers a more accurate measurement of health status and thus should be more powerful to detect treatment effects [McHorney, Haley and Ware (1997), Kosinski *et al.* (2003)]. The benefit of combining sequential analysis and IRT methodologies using mixed Rasch models for binary items has already been studied in the context of non-comparative phase II trials and seems promising [Sébille and Mesbah (2005)]. The joint use of IRT modelling and sequential analysis is extended to comparative phase II and phase III trials using the TT. The test statistics (score statistics and Fisher information for the parameter of interest) used for sequential monitoring of QoL endpoints are derived and studied through simulations.

1.2 IRT MODELS

Item Response Theory (IRT) or more precisely parametric IRT, which was first mostly developed in educational testing, takes into account the multiplicity and categorical nature of the items by introducing an underlying response model [Ficher and Molenaar (1995)] relating those items to a latent parameter interpreted as the true individual QoL. In this framework, the probability of response of a patient on an item depends upon two different kinds of parameters: the "ability level" of the person (which reflects his/her current QoL) and a set of parameters characterizing each item. The basic assumption for IRT models is the unidimensionality property stating that the responses to the items of a questionnaire are influenced by one underlying concept (e.g., QoL) often called latent trait and noted θ. In other words, the person's ability or the person's QoL should be the only variable affecting individual item response. Another important assumption of IRT models, which is closely related to the former, is the concept of local independence meaning that items should be conditionally independent given the latent trait θ. Hence, the joint probability of a response pattern given the latent trait θ can be written as a product of marginal probabilities. Let X_{ij} be the answer for subject i to item j and let θ_i be the unobserved latent variable for subject i ($i = 1, ..., N$; $j = 1, ..., J$):

$$P(X_{i1} = x_{i1}, ..., X_{iJ} = x_{iJ}/\theta_i) = \prod_{j=1}^{J} P(X_{ij} = x_{ij}/\theta_i) \quad (1.1)$$

A last assumption for IRT models is the monotonicity assumption stating that the item response function $P(X_{ij} > k/\theta_i)$ is a non-decreasing function of θ_i, for all j and all k.

1.2.1 The Rasch Model

For binary items, one of the most commonly used IRT model is the Rasch model, sometimes called the one parameter logistic model [Rasch (1980)]. The Rasch model specifies the conditional probability of a patient's response X_{ij} given the latent variable θ_i and the item parameters β_j:

$$P(X_{ij} = x_{ij}/\theta_i, \beta_j) = f(x_{ij}/\theta_i, \beta_j) = \frac{e^{x_{ij}(\theta_i - \beta_j)}}{e^{\theta_i - \beta_j}} \quad (1.2)$$

where β_j is often called the difficulty parameter for item j ($j = 1, ..., J$). Contrasting with other IRT models, in the Rasch model, a patient's total score, $S_i = \sum_{j=1}^{J}$ is a sufficient statistic for a specific latent trait θ_i.

1.2.2 Estimation of the parameters

Several methods are available for estimating the parameters (the θs and βs) in the Rasch model [Fisher and Molenaar (1995)] including: joint maximum likelihood (JML), conditional maximum likelihood (CML), and marginal maximum likelihood (MML). JML is used when person and item parameters are considered as unknown fixed parameters. However, this method gives asymptotically biased and inconsistent estimates [Haberman (1977)]. The second method CML consists in maximizing the conditional likelihood given the total score in order to obtain the items parameters estimates. The person parameters are then estimated by maximizing the likelihood using the previous items parameters estimates. This method has been shown to give consistent and asymptotically normally distributed estimates of item parameters [Andersen (1970)]. The last method MML is used when the Rasch model is interpreted as a mixed model with θ as a random effect having distribution $h(\theta/\xi)$ with unknown parameters ξ. The distribution $h(.)$ is often assumed to belong to some family distribution (often Gaussian) and its parameters are jointly estimated with the item parameters. As with the CML method, the MML estimators for the item parameters are asymptotically efficient [Thissen (1982)]. Furthermore, since MML does not presume existence of a sufficient statistic (unlike CML), it is applicable to

virtually any type of IRT model.

1.3 SEQUENTIAL ANALYSIS

1.3.1 Traditional Sequential Analysis

Let us assume a two-group parallel design with two treatment groups ($g = 1$ for the control group and $g = 2$ for the experimental treatment group) and that the primary endpoint is QoL at the end of the treatment period which is measured using a QoL questionnaire with J dichotomous items. In the traditional framework of sequential analysis [Wald (1947), Whitehead (1997), Jennison and Turnbull (1999)], QoL is assumed to be observed (not to be a latent variable) in each treatment group and the QoL score S_{ig} is used in place of the true latent trait θ_{ig} $(g = 1, 2)$ at each sequential analysis. In that setting, the observed scores in each group (s_{11}, s_{12},...) and (s_{21}, s_{22},...) are assumed to follow some distribution often assumed to be Gaussian with unknown parameters μ_g $(g = 1, 2)$ and common σ_S. Suppose we are testing the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu$ against the one-sided alternative $H_1 : \mu_2 > \mu_1$. The following parameterization is often used for the measure of treatment difference (parameter of interest) $\phi_S = \frac{\mu_2 - \mu_1}{\sigma_S}$. The log-likelihood, which can be expressed according to both independent samples, and its derivatives can be used to derive the test statistics Z(S) and V(S), both evaluated under the null hypothesis. The test statistic Z(S) is the efficient score for ϕ depending on the observed scores S, and the test statistic V(S) is Fisher's information for ϕ.

More precisely, the test statistics Z(S) and V(S) are given by:

$$Z(S) = \frac{n_1 n_2}{(n_1 + n_2)D}(\overline{s}_2 - \overline{s}_1) \tag{1.3}$$

and

$$V(S) = \frac{n_1 n_2}{(n_1 + n_2)} - \frac{Z^2(S)}{2(n_1 + n_2)} \tag{1.4}$$

in which :

- n_g is the cumulated number of patients (since the beginning of the trial) in group g $(g = 1, 2)$,
- $\overline{s}_g = \frac{\sum_{j=1}^{n_g} s_{gj}}{n_g}$ where s_{gj} denotes the observed scores of patient j in group g,

- D is the maximum likelihood estimate of σ_S under the null hypothesis : $D = \sqrt{\frac{Q}{n_1+n_2} - \left(\frac{R}{n_1+n_2}\right)^2}$ with $Q = \sum_{j=1}^{n_1} s_{1j}^2 + \sum_{j=1}^{n_2} s_{2j}^2$ and $R = \sum_{j=1}^{n_1} s_{1j} + \sum_{j=1}^{n_2} s_{2j}$.

Details of the computations are described at length by Whitehead (1997).

1.3.2 Sequential Analysis based on Rasch models

We shall now be interested in the latent case, i.e., the case where θ_{ig} $(g = 1, 2)$ is unobserved in each treatment group. Let us assume that the latent traits θ_1 and θ_2 are random variables that follow normal distributions $N(\psi_1, \sigma_\theta^2)$ and $N(\psi_2, \sigma_\theta^2)$, respectively and that we are testing: $H_0 : \psi_1 = \psi_2 = \psi$ against $H_1 : \psi_1 > \psi_2$. A reparameterization can be performed so that $\varphi = \frac{\psi_2 - \psi_1}{2}$ be the parameter of interest and the nuisance parameter be made up of $\phi = \frac{\psi_1+\psi_2}{2}$ and $\eta = (\sigma, \beta_1, ..., \beta_J)$ such that $\varphi = 0$ under H_0, $\psi_1 = \phi - \varphi$, and $\psi_2 = \varphi + \phi$. Assuming that $n_1 + n_2 = N$ data have been gathered so far in the two treatment groups, the log-likelihood of φ, ϕ and η can be written as $l(\varphi, \phi, \eta) = l^{(1)}(\psi_1, \sigma_\theta, \beta_1, ..., \beta_J) + l^{(2)}(\psi_2, \sigma_\theta, \beta_1, ..., \beta_J)$. Assuming a Rasch model for patient's items responses, we can write:

$$l^{(g)}(\psi_g, \sigma_\theta, \beta_1, ..., \beta_J) = \sum_{i=1}^{N} \log \left\{ \frac{1}{\sigma_\theta \sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(\theta-\psi_g)^2}{2\sigma_\theta^2}} \prod_{j=1}^{J} \frac{e^{x_{ijg}(\theta-\beta_j)}}{1+e^{\theta-\beta_j}} d\theta \right\}, \; g = 1, 2 \tag{1.5}$$

Let ϕ^* and $\eta^* = (\sigma_\theta^*, \beta_1^*, ..., \beta_J^*)$ be the estimates of ϕ and $\eta = (\sigma_\theta, \beta_1, ..., \beta_J)$ under the assumption that both series of data are drawn from the same distribution. There is no analytical solution for ϕ^* and η^* and numerical integration methods have to be used to estimate these parameters. The identifiability constraint $\sum_{j=1}^{J} \beta_j = 0$ is used.

The test statistics Z(X) and V(X), which were previously noted as Z(S) and V(S), will be depending this time directly on X, the responses to the items. They can be derived in the following way:

$$Z(X) = \frac{\partial l(0, \phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_j^*)}{\partial \varphi} = \frac{\partial l^{(2)}(\phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_j^*)}{\partial \psi_2} - \frac{\partial l^{(1)}(\phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_j^*)}{\partial \psi_1} \tag{1.6}$$

That is, we need to evaluate

$$\sum_{i=1}^{N} \frac{\partial}{\partial \psi_g} \left[\log \left(\int_{-\infty}^{+\infty} h_{\psi_g, \sigma_\theta}(\theta) \prod_{j=1}^{J} f(x_{ijg}/\theta; \beta_j) d\theta \right) \right] \tag{1.7}$$

at $(\phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_J^*)$ for $g = 1, 2$ where $h_{\psi_g, \sigma_\theta}$ is the density of the normal distribution.

The test statistic V(X) can sometimes be approximated under H_0 by:

$$\begin{aligned} V(X) &= -\frac{\partial^2 l(0, \phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_J^*)}{\partial \varphi^2} \\ &= -\frac{\partial^2 l^{(2)}(\phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_J^*)}{\partial \psi_2^2} - \frac{\partial^2 l^{(1)}(\phi^*, \sigma_\theta^*, \beta_1^*, ..., \beta_J^*)}{\partial \psi_1^2} \end{aligned} \quad (1.8)$$

when the two samples are large, of about the same size and when φ is small.

Estimation of the statistics Z(X) and V(X) is done by maximising the marginal likelihood, obtained from integrating out the random effects. Quasi-Newton procedures can be used for instance to maximise the likelihood and adaptive Gaussian quadrature can be used to integrate out the random effects [Pinheiro and Bates (1995)].

1.3.3 The Triangular Test

For the ease of the general presentation of the sequential test we shall use the conventional notations Z and V. The TT uses a sequential plan defined by two perpendicular axes, the horizontal axis corresponds to Fisher's information V, and the vertical axis corresponds to the efficient score Z which represents the benefit as compared with H_0. For a one-sided test, the boundaries of the test, delineate a continuation region (situated between these lines), from the regions of non rejection of H_0 (situated beneath the bottom line) and of rejection of H_0 (situated above the top line). 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 [Siegmund (1979)]. The expressions of the boundaries for a one-sided test are well-known [Sébille and Bellissant (2001)]. At each analysis, the values of the two statistics Z and V are computed and Z is plotted against V, thus forming a sample path as the trial goes on. 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: non rejection of H_0 if the sample path crosses the lower boundary, and rejection of H_0 if it crosses the upper boundary. This test and other types of group sequential tests are implemented in the computer program PEST 4 [MPS Research Unit (2000)] that can be used for the planning, monitoring and analysis of comparative clinical trials.

1.4 SIMULATIONS

1.4.1 Simulations design

The statistical properties of the TT were evaluated with simulated data. We studied the type I error (α), the power ($1-\beta$), and the average sample number (ASN) of patients required to reach a conclusion. A thousand comparative clinical trials were simulated. The latent trait in the control group θ_{i1} was assumed to follow a normal distribution with mean λ_1 and variance $\sigma^2 = 1$ and the latent trait in the experimental group θ_{i2} was assumed to follow a normal distribution with mean $\lambda_2 = \lambda_1 + d$ and same variance. The trial involved the comparison of the two hypotheses: $H_0 : d = 0$ against $H_1 : d > 0$.

We first assumed that the items under consideration formed part of a calibrated item bank, meaning that items parameters were assumed to be known [Holman *et al.* (2003)]. We also investigated the more extreme case where all items parameters are assumed to be totally unknown and have therefore to be estimated at each sequential analysis. For both cases, the items parameters were uniformly distributed in the interval [-2, 2] with $\sum_{j=1}^{J} \beta_j = 0$.

The traditional method consisted in using the observed QoL scores, S, given by the sum of the responses to the items, which were assumed to follow a normal distribution. The Z(S) and V(S) statistics were computed within the well-known framework of normally distributed endpoints [20].

We compared the use of Rasch modelling methods with QoL scores methods. To evaluate the effect of the number of items used for measuring QoL, we investigated QoL questionnaires with 5 or 10 items. Moreover, different expected effect sizes (noted ES equal to $\frac{\lambda_2-\lambda_1}{\sigma} = d$) ranging from small (0.4) to large (0.8) were investigated. The sequential analyses were performed every 40 included patients and $\alpha = \beta = 0.05$ for all simulations.

The simulations have been perfomed using a C++ program, and the data have been analysed with the SAS software [Hardouin and Mesbah (2007)].

1.4.2 Results

Table 1.1 shows the type I error and power for the TT for different values of the effect size and of the number of items using the method based on QoL scores or the Rasch modelling method assuming either known or unknown items parameters values. The significance level was usually close to the target value of 0.05 for the QoL scores method and the Rasch modelling method assuming known items parameters values. However, the significance level was always

Table 1.1: Type I error and Power for the Triangular Test (TT) using the method based on QoL scores or the Rasch model for different values of the effect size and of the number of items (nominal $\alpha = \beta = 0.05$, 1000 simulations)

Effect size	Number of items	Type I error			Power		
		QoL scores	Rasch model		QoL scores	Rasch model	
			β known	β unknown		β known	β unknown
0.4	5	0.027	0.039	0.058	0.758	0.951	0.926
0.4	10	0.045	0.044	0.082	0.852	0.952	0.926
0.5	5	0.039	0.048	0.077	0.736	0.944	0.908
0.5	10	0.057	0.064	0.088	0.838	0.951	0.931
0.6	5	0.045	0.056	0.072	0.736	0.934	0.907
0.6	10	0.052	0.057	0.083	0.846	0.952	0.934
0.7	5	0.044	0.046	0.076	0.743	0.938	0.912
0.7	10	0.054	0.049	0.079	0.844	0.947	0.932
0.8	5	0.049	0.041	0.069	0.741	0.943	0.924
0.8	10	0.055	0.049	0.080	0.836	0.949	0.941

higher than the target value of 0.05 for the Rasch modelling method assuming unknown items parameters values for all effect sizes and number of items considered. The TT was quite close to the nominal power of 0.95 when the Rasch modelling method assuming known items parameters values was used, and a little lower than expected when unknown items parameters values were assumed. However, the TT was notably underpowered when the QoL scores method was used. Indeed, for the QoL scores method, as compared with the target power value of 0.95, there were decreases in power of approximately 22% and 11% with 5 and 10 items, respectively. By contrast, for the Rasch modelling method assuming unknown items parameters values, the decrease in power was of about only 4% and 2% with 5 and 10 items, respectively.

Table 1.2 shows the ASN of the number of patients required to reach a conclusion under H_0 and H_1 for the TT for different values of the effect size and of the number of items using the method based on QoL scores or the Rasch modelling method assuming either known or unknown items parameters values. We also computed for comparison purposes the number of patients required by the single-stage design (SSD) and the approximate ASN for the TT computed with PEST 4 when a normally distributed endpoint is assumed when planning the trial. As expected, the ASNs all decreased as the expected effect sizes increased whatever the method used. The ASNs under H_0 and H_1 were always smaller for all sequential procedures based either on QoL scores or Rasch modelling methods than the sample size required by the SSD for whatever values of effect size or number of items considered. The decreases in the ASNs under H_0 and H_1 were usually more marked when the Rasch modelling methods were used, assuming either known or unknown items parameters values, as compared with the methods based on QoL scores. Indeed, under H_0 (H_1) as compared

Table 1.2: 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 Triangular Test (TT) using the method based on QoL scores or the Rasch model for different values of the effect size and of the number of items (nominal $\alpha = \beta = 0.05$, 1000 simulations)

Effect size	Number of items	SSD	TT*	QoL scores	Rasch model	
					β known	β unknown
			H_0/H_1	H_0/H_1	H_0/H_1	H_0/H_1
0.4	5	271	155 / 155	140 / 178	140 / 148	135 / 145
0.4	10	271	155 / 155	141 / 167	117 / 122	114 / 119
0.5	5	174	103 / 103	104 / 128	102 / 103	102 / 92
0.5	10	174	103 / 103	103 / 121	84 / 85	83 / 84
0.6	5	121	74 / 74	76 / 95	77 / 76	77 / 77
0.6	10	121	74 / 74	76 / 91	62 / 63	64 / 63
0.7	5	89	57 / 57	60 / 72	61 / 60	63 / 60
0.7	10	89	57 / 57	60 / 70	51 / 51	53 / 52
0.8	5	68	46 / 46	50 / 58	51 / 51	52 / 52
0.8	10	68	46 / 46	50 / 56	45 / 45	47 / 45

*: Approximate ASN for the TT for a normally distributed endpoint.

with the SSD, there were decreases of approximately 37% (25%) and 41% (42% in sample sizes for the QoL scores method and the Rasch modelling methods, respectively.

1.5 DISCUSSION - CONCLUSION

We evaluated the benefit of combining sequential analysis and IRT methodologies in the context of phase II or phase III comparative clinical trials using QoL endpoints. We studied and compared the statistical properties of a group sequential method, the TT, using either mixed Rasch models assuming either known or unknown items parameters values or the traditional method based on QoL scores. Simulation studies showed that: (i) the type I error a was correctly maintained for the QoL scores method and the Rasch modelling method assuming known items parameters values but was always higher than expected for the Rasch modelling method assuming unknown items parameters values, (ii) the power of the TT was correctly maintained for the Rasch modelling method assuming known items parameters values and a little lower than expected when items parameters were assumed to be unknown, but the TT was particularly underpowered for the QoL scores method, (iii) as expected using group sequential analysis, all methods allowed substantial reductions in ASNs as compared with the SSD, the largest reduction being observed with the Rasch modelling methods.

The different results that were obtained using the mixed Rasch models assuming either known or unknown items parameters values or the method based on QoL scores might be partly explained by looking at the distributions of the test statistics Z(S), V(S), Z(X), and V(X). According to asymptotic distributional results, we might expect the sequences of test statistics ($Z_1(S)$, $Z_2(S)$,..., $Z_K(S)$) and ($Z_1(X)$, $Z_2(X)$,..., $Z_K(X)$) to be multivariate normal with: $Z_k(S) \sim N(ES * V_k(S), V_k(S))$ and $Z_k(X) \sim N(ES * V_k(X), V_k(X))$, respectively, where ES denotes the effect size, for $k = 1, 2, ..., K$ analyses [Whitehead (1997), Jennison and Turnbull (1999)]. Table 1.3 shows the distribution of the standardized test statistics under H_0 and H_1 (effect size equal to 0.5) that were estimated using the method based on QoL scores or the Rasch models assuming either known or unknown items parameters values. The estimation of the test statistics were performed at the second sequential analysis corresponding to a sample size of 80 patients. The normality assumption was not rejected using a Kolmogorov-Smirnov test, whatever the method used. Under H_0 or H_1, the null hypothesis of unit standard deviation (SD) was rejected when the estimation was performed with the mixed Rasch model assuming unknown items parameters values, the estimated SD being larger than expected. This feature might be to some extent responsible of the inflation of the type I error a under H_0 and might also partly explain the bit of under powering of the TT that was observed under most H_1 hypotheses. Under H_1, the null hypothesis of 0 mean was rejected when the estimation was performed with the QoL scores method, the estimated mean value being lower than expected. This might explain why the TT was notably underpowered using the QoL scores method.

Another important aspect is also to be noted for the mixed Rasch model assuming unknown items parameters values. The use of this model corresponds to a rather extreme case where no information is assumed to be known about the items parameters. This can be the case if no data have ever been collected using the corresponding QoL questionnaire, which is rarely the case. Otherwise, one could use data from another study using that specific QoL questionnaire to estimate the items parameters and then use these estimates in the Rasch model, since the items parameters are assumed to be parameters related only to the questionnaire and are therefore supposed to be invariant from one study to another (using the same QoL questionnaire). In our simulation study and in the example using the data from the phase III oncology trial, the items parameters were estimated at each sequential analysis, that is on 40, 80, 120, ... patients since the group sequential analyses were performed every 40 patients. It is very likely that the amount of available data at each sequential analysis might be quite insufficient to satisfactorily estimate the item difficulty parameters, especially when estimating 5 or 10 items with only 40 patients. The

Table 1.3: Distribution of the standardized test statistics estimated using the method based on QoL scores or the Rasch model for different values of the number of items and for an effect size equal to 0.5, assuming that the vector of items parameters values b is either known or unknown (nominal $\alpha = \beta = 0.05$, 1000 simulations)

Number of items	H_0			H_1		
	QoL scores	Rasch model		QoL scores	Rasch model	
		β known	β unknown		β known	β unknown
	Z'(S)	Z'(X)	Z'(X)	Z'(S)	Z'(X)	Z'(X)
5	-0.034	-0.005	-0.028	-0.654*	-0.009	-0.006
	(0.995)	(0.972)	(1.090)**	(1.014)	(0.978)	(1.086)**
10	-0.037	-0.003	-0.016	-0.423*	0.007	0.029
	(0.995)	(1.017)	(1.143)**	(1.009)	(0.996)	(1.131)**

Z'(S) and Z'(X) are the standardized test statistics for the method based on QoL scores and the Rasch model, respectively : $Z'(S) = \frac{Z(S)-ES.V}{\sqrt{V}}$ and $Z'(X) = \frac{Z(X)-ES.V}{\sqrt{V}}$ where ES is the effect size. Data are means (SD).
$* : p < 0.001$ for testing the mean equal to 0
$** : p < 0.05$ for testing the standard deviation equal to 1

simulations were also performed using 80 patients for the first sequential analysis to estimate the items parameters and 40 more patients at each subsequent sequential analysis and this resulted in a type I error closer to the target value of 0.05 and higher power (data not shown). However, it has to be mentioned that such a feature might not be interesting for larger effect sizes (over 0.6) because the benefit in terms of ASNs offered by sequential analyses might then be overwhelmed by the fact that it will not be possible to stop the study before 80 patients have been included.

We obtained conflicting results when using the QoL scores method and the mixed Rasch model assuming unknown items parameters values on the data from the phase III oncology trial. As the non-sequential analysis suggested from the results obtained for the physical functioning scale at the end of the first cycle of treatment, we might expect not to be under the null hypothesis of no treatment effect. The lack of power of the TT which was displayed in the simulations when using the method based on QoL scores might be responsible for not rejecting the null hypothesis and this might reflect that a type II error has occurred. The Rasch modelling method assuming unknown items parameters values (with items parameters being estimated at each sequential analysis) allowed rejection of the null hypothesis but gave estimates of the test statistic V(X) which seemed to be importantly underestimated. This might be explained by different aspects: (i) the fit of the Rasch model seemed to be poor especially for item 5, (ii) the successive estimates of the SD of the latent trait q at each sequential analysis were large (of about 5) as well as the standard errors of the estimated items parameters reflecting an important lack of precision in

the parameters estimates. Despite these difficulties, the method seemed to be able to reach the most likely correct conclusion (rejection of the null hypothesis).

Other types of investigations on incorporating IRT methodologies in sequential clinical trials could also be interesting to perform such as: evaluating the impact on the statistical properties of the sequential tests of the amount of missing data (often encountered in practice and not investigated in our study) and missing data mechanisms (missing completely at random, missing at random, non ignorable missing data). In addition, other group sequential methods could also be investigated such as spending functions [Lan and De Mets (1983)], and Bayesian sequential methods [Grossman *et al.* (1994)] for instance. Finally, we only worked on binary items and polytomous items more frequently appear in health-related QoL questionnaire used in clinical trial practice. Other IRT models such as the Partial Credit Model or the Rating Scale Model [Fisher and Molenaar (1995)] would certainly be more appropriate in this context.

Item response theory usually provides more accurate assessment of health status as compared with QoL scores method [McHorney, Haley and Ware (1997), Kosinsky *et al.* (2003)]. The use of IRT methods in the context of sequential analysis of QoL endpoints provides a more powerful method to detect therapeutic effects than the traditional method based on QoL scores. Finally, there are a number of challenges for medical statisticians using IRT that may be worth to mention: IRT was originally developed in educational research using samples of thousands or even ten thousands. Such large sample sizes are very rarely (almost never) attained in medical research where medical interventions are often assessed using less than 200 patients. The problem is even more crucial in the sequential analysis framework where the first interim analysis is often performed on fewer patients. Moreover, IRT and associated estimation procedures are conceptually more difficult than the QoL scores method often used in medical research. Perhaps one of the biggest challenges for medical statisticians will be to explain these methods well enough so that clinical researchers will accept them and use them. As in all clinical research but maybe even more in this context, there is a real need for good communication and collaboration between clinicians and statisticians.

1.6 REFERENCES

1. Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators, *Journal of the Royal Statistical Society series B*, **32**, 283–301.

2. Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size, *Annals of Mathematical. Statistics*, **31**, 165–197.

3. Cella, D. F, Bonomi, A. E. (1995). Measuring quality of life: 1995 update, *Oncology*, **9**, 47–60.

4. Fairclough, D. L. (2002). *Design and analysis of quality of life studies in clinical trials*, Chapman & Hall/CRC: Boca Raton.

5. Fisher, G. H. and Molenaar, I. W. (1995). *Rasch Models, Foundations, Recent Developments, and Applications*, Springer-Verlag: New-York.

6. Grossman, J., Parmar, M. K., Spiegelhalter, D.J., Freedman, L.S. (1994). A unified method for monitoring and analysing controlled trials, *Statistics in Medicine*, **13**, 1815–1826.

7. Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models, *Annals of Statistics*, **5**, 815–841.

8. Hardouin, J. B. and Mesbah M. (2007). The SAS macro-program %AnaQol to estimate the parameters of Item Response Theory Models, *Communications in Statistics - Simulation and Computation*, **36**, in press.

9. Holman, R., Glas C. A., de Haan, R. J. (2003). Power analysis in randomized clinical trials based on item response theory, *Controlled Clinical Trials*, **24**, 390–410.

10. Holman, R., Lindeboom, R., Glas, C. A. W., Vermeulen, M., de Haan, R. J. (2003). Constructing an item bank using item response theory: the AMC linear disability score project, *Health Services and Outcomes Research Methodology*, **4**, 19–33.

11. Jennison, C. and Turnbull, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC: Boca Raton.

12. Kosinski, M., Bjorner, J. B., Ware, J. E. Jr, Batenhorst, A. and Cady, R. K. (2003). The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: a re-analysis of three clinical trials, *Quality of Life Research*, **12**, 903–912.

13. Lan, K. K. G., De Mets, D. L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika*, **70**, 659–663.

14. McHorney, C. A., Haley, S. M. and Ware, J. E. Jr. (1997). 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*, **50**, 451–461.

15. MPS Research Unit. (2000). *PEST 4: operating manual*, The University of Reading, Reading.

16. O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials, *Biometrics*, **35**, 549–556.

17. Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model, *Journal of Computational and Graphical Statistics*, **4**, 12–35.

18. Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika*, **64**, 191–199.

19. Rasch, G. (1980). *Probabilistisc models for some intelligence and attainment tests.* Copenhagen, D.K.:Nielsen & Lydiche. Expanded edition, Chicago: The University of Chicago Press.

20. Sébille, V., Bellissant E. (2001). Comparison of the two-sided single triangular test to the double triangular test, *Controlled Clinical Trials*, **22**, 503–514.

21. Sébille, V. and Mesbah, M. (2005). Sequential Analysis of Quality of Life Rasch Measurements. In *Probability, Statistics and Modelling in Public Health* (Ed. M. Nikulin, D. Commenges, C. Huber), Springer. In press.

22. Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems, *Advances in Applied Probability*, **11**, 701-719.

23. Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model, *Psychometrika*, **47**, 175–186.

24. Wald, A. (1947). *Sequential Analysis*, Wiley: New York.

25. Whitehead, J. and Jones, D. R. (1979). The analysis of sequential clinical trials, *Biometrika*, **66**, 443–452.

26. Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions, *Biometrics*, **39**, 227–236.

27. Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials, revised 2nd edition*, Wiley: Chichester.