Special Issue Paper

Statistics in Medicine

Received 27 October 2010, Accepted 4 August 2011 Published online 8 November 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4387

Towards power and sample size calculations for the comparison of two groups of patients with item response theory models

Jean-Benoit Hardouin,[a*†] Sarah Amri,[a,b] Mohand-Larbi Feddag[a] and Véronique Sébille[a]

Evaluation of patient-reported outcomes (PRO) is increasingly performed in health sciences. PRO differs from other measurements because such patient characteristics cannot be directly observed. Item response theory (IRT) is an attractive way for PRO analysis. However, in the framework of IRT, sample size justification is rarely provided or ignores the fact that PRO measures are latent variables with the use of formulas developed for observed variables. It might therefore be inappropriate and might provide inadequately sized studies. The objective was to develop valid sample size methodology for the comparison of PRO in two groups of patients using IRT. The proposed approach takes into account questionnaire's items parameters, the difference of the latent variables means, and its variance whose derivation is approximated using Cramer–Rao bound (CRB). We also computed the associated power. We realized a simulation study taking into account sample size, number of items, and value of the group effect. We compared power obtained from CRB with the one obtained from simulations (SIM) and with the power based on observed variables (OBS). For a given sample size, powers using CRB and SIM were similar and always lower than OBS. We observed a strong impact of the number of items for CRB and SIM, the power increasing with the questionnaire's length but not for OBS. In the context of latent variables, it seems important to use an adapted sample size formula because the formula developed for observed variables seems to be inadequate and leads to an underestimated study size. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: item response theory; sample size; Rasch model; simulation; power

1. Introduction

The evaluation of perceived health measures or more generally 'patient-reported outcomes' (PRO) is increasingly performed in clinical and epidemiological research, especially in chronic conditions. These measures often include health-related quality of life (QoL), depression, or pain outcomes for instance. PRO differs from other measurements because such patient characteristics cannot be directly observed and measured and are usually evaluated using self-assessment questionnaires, which consist of a set of items whose responses provided by the patients are combined to give scores. Two main types of analytic strategies are used for such data: the so-called classical test theory (CTT) and models coming from item response theory (IRT). CTT relies on the observed scores (possibly weighted sum of patients' items responses) that are assumed to provide a good representation of a 'true' score, whereas IRT relies on an underlying response model relating the items responses to a latent parameter, often called latent trait, and usually assumed to follow a gaussian distribution, interpreted as the true individual QoL, for instance.

Despite the widespread use of PRO in clinical research, the design and planning of the corresponding studies are often faced with some important methodological issues, such as the statistical calculation of

[a] *UPRES EA 4275 Biostatistics, Clinical Research and Subjective Measures in Health Science, University of Nantes, France*

[b] *ENS Cachan, Campus de Ker-Lann, Bruz, France*

[*] *Correspondence to: Jean-Benoit Hardouin, EA 4275 Biostatistics, Clinical Research and Subjective Measures in Health Sciences, Faculty of Pharmaceutical Sciences, University of Nantes, 1 Rue Gaston Veil, BP 53508, 44035 Nantes Cedex 1, France.*

[†] *E-mail: jean-benoit.hardouin@univ-nantes.fr*

sample size. As a matter of fact, the justification of study size remains hardly ever provided in the framework of self-reported perceived health measures. Furthermore, it has been stressed that many studies might not be adequately powered to determine clinically important changes in QoL or more generally symptom control [1, 2]. Specific sample size methodology is importantly needed for clinical research including PRO to avoid inadequately sized studies that may lead to erroneous and uninformative conclusions and may expose patients to inappropriate medical strategies. One of the main issues in study design including PRO arises from the type of endpoint being measured and from its major attribute: the fact that it is an unobserved latent variable and that it should be managed and analyzed as such with appropriate modeling strategies, particularly when designing a study as soon as planning phase, for reliable sample size calculations and further analyses. Indeed, to date, most sample size calculations, if any, come from CTT that often assumes the normality of scores, which is rarely encountered in practice and may lead to inadequately sized studies. Some authors have also suggested the use of nonparametric methods [3, 4] that also rely on some assumptions and are seldom used in practice.

Methods coming from modern measurement theory, such as IRT models, might provide a powerful framework to build and reduce PRO instruments and analyze such data in a more efficient and reliable manner and should provide valid measures of QoL, anxiety, or pain for instance [5]. Indeed, models coming from IRT are more and more used for the construction, validation and reduction of questionnaires [6, 7], and for creating item banks [8]. Hence, a lot of PRO instruments are found to be well adapted to IRT modeling either because of the way they were developed using such IRT-based strategies (for example, [9]) or because of their desired psychometric properties (for example, [10, 11]). Moreover, IRT is sometimes used to obtain short version of the questionnaires (for example, [12]), notably with the use of computerized adaptive testing [13–16], and is sometimes presented as a solution to have an appropriate management of possible floor and ceiling effects. IRT also allows for the comparison of scores between different instruments and obtaining interval measure for the latent trait [17]. It will be appropriate to compare IRT score means, for instance by contrast, with CTT that only produces an ordinal measure of the latent trait.

Furthermore, good methodological standards recommend that methods used for sample size planning and for statistical analysis should be based on similar grounds. Hence, if IRT models best describe data coming from PRO instruments, they should be taken into account right away during the planning of the study and sample size calculations. However, in the framework of IRT modeling, sample size determination is often either not performed at all or relies on CTT and on expected mean scores, thus ignoring the fact that these specific measures are in fact latent variables [5]. Only very recent developments concern the question of sample size or power in IRT [18], particularly for the validation of a questionnaire by an IRT model [19].

An alternative method could be to directly apply the classical sample size formula developed for normally distributed endpoints [20] on the latent variable measured by an IRT model. This approach could be attractive, particularly when the latent trait is assumed to be a random variable following a Gaussian distribution. However, because the usual formula relies on the so-called manifest variables, that is, observable variable that can be measured directly, and not on latent variables, it might be inadequate. Indeed, previous work [18] has shown that using this formula for latent variables is not correct as it leads to underestimation of the required sample size, and that adapted sample size calculations have to be developed in the context of latent variables and IRT modeling.

The main objective of this study is to provide valid sample size methodology in the framework of the comparison of PRO in two groups of patients in cross-sectional studies using IRT modeling strategies. More specifically, we have focused on the Rasch model as a special case of an IRT model to illustrate the proposed approach. In this context, we develop and validate a theoretical methodological approach using simulation studies. We compare the required sample size computed with this approach, adapted to the context of the Rasch model, with the required sample size provided by the classical formula used in the context of manifest variables.

2. Methods

2.1. Item response theory models

Some of the commonly used IRT models, particularly in health sciences, are the Rasch model for binary responses [21, 22] and the rating scale model or the partial credit model for multiple (>2) response options [23, 24]. We will consider the Rasch model and discuss extensions of the results to other IRT models. IRT is based on three assumptions, which are often desirable notably during the construction and validation steps of PRO instruments: (i) unidimensionality stating that one latent trait influences the responses to all the items; (ii) local independence meaning that for a given individual, the responses to the items are independent; and (iii) monotonicity stating that the probability to have a positive response to a given item does not decrease with the latent variable. The Rasch model can be implemented as a random effects model to reflect the fact that the sample of patients is assumed to be representative of a more general population. Hence, the latent variable of interest (Θ) is considered as a random variable.

Assume that N patients have answered a questionnaire containing J dichotomous items. Let X_{nj} be the random variable representing the response of patient n to item j with realization x_{nj} and θ_n the realization of the latent trait for this patient. For each patient ($n = 1, ..., N$), the response probability of the jth item is

$$P(X_{nj} = x_{nj}|\theta_n, \delta_j) = \frac{\exp\left\{x_{nj}(\theta_n - \delta_j)\right\}}{1 + \exp(\theta_n - \delta_j)} \tag{1}$$

where δ_j represents the difficulty of item j.

The variables $\theta_1, \theta_2, ... \theta_N$ are mutually independent with a common underlying distribution, which is often assumed to be Gaussian.

2.2. *Estimation of the parameters*

With the use of the local independence assumption, the marginal likelihood can be written down and the parameters of the latent trait can be jointly estimated with the item parameters by marginal maximum likelihood (MML) estimation obtained from integrating out the random effects [21].

$$L(\sigma^2, \delta_1, ..., \delta_J|x) = \prod_{n=1}^{N} \int \prod_{j=1}^{J} \frac{\exp\left\{x_{nj}(\theta - \delta_j)\right\}}{1 + \exp(\theta - \delta_j)} G(\theta|\sigma^2)\mathrm{d}\theta \tag{2}$$

where $G(.|\sigma^2)$ is the Gaussian distribution function with mean $\mu = 0$ and variance σ^2. The constraint $\mu = 0$ is an identifiability constraint, allowing the estimation of the other parameters conditionally on it [21].

Quasi-Newton algorithm is often used to maximize the likelihood along with adaptive Gauss–Hermite quadrature to integrate out the random effects [25]. The MML estimators that are obtained are asymptotically efficient [21, 26].

2.3. *Sample size calculations for item response theory analysis*

Suppose we plan to conduct a cross-sectional study for the comparison of two groups of patients. Let N_0 and N_1 be the expected sample sizes in each group and $N = N_0 + N_1$. Let Θ be the latent trait with normal distributions $N(-\frac{N_1}{N}\gamma, \sigma^2)$ and $N(\frac{N_0}{N}\gamma, \sigma^2)$ in the first (coded 0) and second (coded 1) groups, respectively. With this parametrization, $\gamma > 0$ represents the difference between the mean values of the latent traits in the two groups. The global mean value of the latent trait among all the N patients is 0 (corresponding to the identifiability constraint, the global mean being the average of the mean of each group weighted by the size of the corresponding group : $\mu = \frac{1}{N_0+N_1}\left(N_0 \times \frac{-N_1}{N}\gamma + N_1 \times \frac{N_0}{N}\gamma\right) = 0$), and σ^2 is the variance of the latent trait (considered to be equal in the two groups). This choice allows considering the same values of the difficulty parameters as the ones estimated during validation step, where these parameters are estimated on the whole validation sample. We shall assume that the study involves the comparison of the two hypotheses: $\mathrm{H}_0 : \gamma = 0$ against $\mathrm{H}_1 : \gamma \neq 0$. Sample size determination is closely related to the Wald test of group effect based on an estimate Γ of γ and on its standard error. The determination of the latter should encompass parameters related to the items of the PRO instrument, reflecting its psychometric properties, as well as the uncertainty related to the estimation of the latent trait according to the IRT model. The derivation of an analytical formula for the standard error of Γ can use Fisher's information obtained using the second derivative of the marginal likelihood and Cramer–Rao's boundary property, providing a lower boundary for the variance. Here, we shall assume that Γ is an unbiased estimator of γ, normally distributed with mean γ and variance var(Γ). Because we are planning a study, we shall make some assumptions and let item parameters and the variance σ^2 of the latent trait be set to some predetermined values (they will be set to fixed values in the simulation study).

2.4. *Fisher's information and the Cramer–Rao bound*

Let Φ be an unknown parameter that can be estimated from data x, distributed according to some probability density function $\mathrm{f}(x;\phi)$. It can be shown that the variance of any unbiased estimator $\hat{\phi}$ of ϕ is bounded by the inverse of Fisher's information $I(\phi)$: $\mathrm{var}(\hat{\phi}) \geqslant 1/I(\phi)$. This is known as the Cramer–Rao inequality, and the number $1/I(\phi)$ is known as the Cramer–Rao lower bound. Fisher's information is defined by the following:

$$I(\phi) = -E\left[\frac{\partial^2 l(\phi|x)}{\partial\phi^2}\right] \tag{3}$$

where $l(x;\phi)$ is the natural logarithm of the likelihood function.

2.5. *Calculation of Fisher's information for the Rasch model*

Because the latent traits follow normal distributions $N(-\frac{N_1}{N}\gamma,\sigma^2)$ and $N(\frac{N_0}{N}\gamma,\sigma^2)$ in the first (coded 0) and second (coded 1) groups with expected sample sizes N_0 and N_1, respectively, we can write the marginal maximum likelihood as follows:

$$L(\sigma^2,\delta_1,\ldots,\delta_J,\gamma|x) = \prod_{g=0}^{1}\prod_{n=1}^{N_g}\int\prod_{j=1}^{J}\frac{\exp(x_{nj}(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j))}{1+\exp(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j)}G(\theta|\sigma^2)\mathrm{d}\theta \tag{4}$$

where $(-1)^{1-g}N_{1-g}\gamma/N$ is a simplified expression of the mean for the latent trait in the gth group ($g = 0, 1$).

During the planning of the study, we assume the values of the parameters to be fixed to some hypothesized values, so we will consider that the parameters $\delta_1,\ldots,\delta_J$ and σ^2 are known. We can then write the log-likelihood defined in Equation (4) as follows:

$$l = l(\gamma|x,\sigma^2,\delta_1,\ldots,\delta_J) = \log\prod_{g=0}^{1}\prod_{n=1}^{N_g}\int\prod_{j=1}^{J}\frac{\exp(x_{nj}(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j))}{1+\exp(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j)}G(\theta|\sigma^2)\mathrm{d}\theta \tag{5}$$

We propose to use the Cramer–Rao lower bound and Equation (3) in order to provide an approximation of the estimate of the variance of γ:

$$\hat{\mathrm{var}}(\gamma) \approx -\left[E\left[\frac{\partial^2 l}{\partial\gamma^2}\right]\right]^{-1} \tag{6}$$

The analytic development of this quantity is defined in Appendix A.

Most of the necessary elements for the calculation of the second derivative of the log-likelihood (thus Fisher's information) are now obtained except one unknown parameter, the binary expected patient's responses, x_{nj} ($n = 1,\ldots,N$ and $j = 1,\ldots,J$). During the planning of the study, the patient's responses always are unknown. As a consequence, we determine a set of expected responses, conditionally on all the other parameters fixed to their expected values: number of patients in each group (N_0 and N_1), number of items (J), group effect (γ), variance of the latent trait (σ^2), and difficulty parameters of the items ($\delta_j,\ j = 1\ldots J$).

First, we determine all the 2^J possible binary response pattern $x^{(p)} = (x_1^{(p)}\ldots x_j^{(p)}\ldots x_J^{(p)}),\ \forall p = 1\ldots 2^J,\ j = 1\ldots J,\ x_j^{(p)} = 0, 1$, and for each of them, we compute two probabilities π_{p0} (for the pth pattern in group 0) and π_{p1} (for the pth pattern in group 1) using the marginal probability to observe the corresponding responses $x_j^{(p)},\ j = 1\ldots j$ (Equation (1)) and the local independence assumption:

$$\pi_{pg} = \int\prod_{j=1}^{J}\frac{\exp\left(x_j^{(p)}\left(\theta-\delta_j\right)\right)}{1+\exp\left(\theta-\delta_j\right)}G(\theta|\mu_g = (-1)^{1-g}N_{1-g}\gamma/N,\sigma^2)\mathrm{d}\theta \tag{7}$$

We evaluate these probabilities using Gauss–Hermite quadratures.

We then determine the expected frequencies n_{pg} of each pattern p in each group g in the following ways:

- We realize a first evaluation of n_{pg} using $n^*_{pg} = \text{floor}(N_g \times \pi_{pg})$ with floor(x) the integer n such that $n \leqslant x < n+1$.
- We compute the numbers of unaffected frequencies $N^*_g = N_g - \sum_p n^*_{pg}$ $(g = 0, 1)$.
- We compute the residual probabilities $\pi^*_{pg} = \pi_{pg} - n^*_{pg}/N_g$.
- We distribute the unaffected frequencies among all the N^*_g patterns having the greatest values of the residual probabilities π^*_{pg} (for each of them, the final frequency is $n_{pg} = n^*_{pg} + 1$, and for the others patterns, we use $n_{pg} = n^*_{pg}$).

The association between each response pattern and its corresponding frequency allows obtaining an expected dataset that can be used to determine the Cramer–Rao bound (CRB).

2.6. *Estimation of the power of the test*

We can test the group effect using a Wald test. The null hypothesis is $H_0 : \gamma = 0$ against the alternative hypothesis $H_1 : \gamma \neq 0$. The statistic of this test is $\frac{\Gamma}{\sqrt{\text{var}(\Gamma)}} \sim N(0,1)$ under H_0. We reject the null hypothesis at level α if $\frac{|\hat{\gamma}|}{\sqrt{\hat{\text{var}}(\hat{\gamma})}} > z_{1-\alpha/2}$ with $z_{1-\alpha/2}$ being the quantiles of the standard normal distribution.

We can evaluate the expected power $1-\hat{\beta}_{\text{CR}}$ based on the CRB associated to this test in the following way:

$$1-\hat{\beta}_{\text{CR}} = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\hat{\gamma}}{\sqrt{\hat{\text{var}}(\hat{\gamma})}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\hat{\gamma}}{\sqrt{\hat{\text{var}}(\hat{\gamma})}}\right) \tag{8}$$

under the alternative hypothesis, with Φ being the cumulative standard normal distribution function. Because γ is estimated by MML, its estimate is unbiased and $\hat{\gamma}$ was set to γ in Equation (8). We evaluate $\hat{\text{var}}(\hat{\gamma})$ using the CRB.

If we assume that the groups have been coded such that the γ parameter will take a positive value, then the second term of the Equation (8) is negligible and so

$$1-\hat{\beta}_{\text{CR}} \approx 1 - \Phi\left(z_{1-\alpha/2} - \frac{\gamma}{\sqrt{\hat{\text{var}}(\gamma)}}\right) \tag{9}$$

This approximation is in accordance with the traditional way to predict the power or to compute a required sample size.

2.7. *Simulation study*

We realize a simulation study to validate the obtained formula for the variance of the group effect and to estimate the power of the test. We simulate data using a Rasch model in two groups, where we drew the latent trait from a Gaussian distribution with mean $-N_1\gamma/N$ and $N_0\gamma/N$ in groups 0 and 1, respectively, and variance equal to 1.

The parameters of this simulation study are the following:

- The number of individuals per group $(N_0 = N_1)$: 50, 100, 200, 300, and 500
- The number of items (J): 5 and 10
- The difficulty parameters of the items: $(-1, -0.5, 0, 0.5, 1)$ for $J = 5$ and $(-2, -1.5, -1, -0.5, 0, 0, 0.5, 1, 1.5, 2)$ for $J = 10$
- The group effect (γ): 0, 0.2, 0.5, and 0.8

This constituted 40 scenarios, and we replicate each of them 1000 times. These scenarios reflect the range of sample sizes and the number of items often encountered in clinical and epidemiological research.

In each replication, the group effect has been estimated by a Rasch model including a group effect, with difficulty parameters and variance of the latent trait assumed to be known and fixed to the expected planning values.

We performed a Wald test of the group effect. An estimate of the type I error obtained by simulation corresponded to the rate of rejection of the null hypothesis at $\alpha = 5\%$ when $\gamma = 0$. An estimate of the power obtained by simulation $(1 - \hat{\beta}_S)$ corresponded to the rate of rejection of the null hypothesis at $\alpha = 5\%$ when $\gamma \neq 0$.

For each scenario, the expected power $(1 - \hat{\beta}_C)$ can also be obtained using the classical formula for manifest variables [20, p. 46, equation 3.6]

$$1 - \hat{\beta}_C = \Phi\left(\sqrt{\frac{rN_0 \times \hat{\gamma}^2}{(r+1) \times \hat{\sigma}^2}} - z_{1-\alpha/2}\right) \tag{10}$$

with $r = N_1/N_0$ fixed to 1 in the simulations. The values of $\hat{\gamma}$ and $\hat{\sigma}^2$ are fixed to their expected planning values γ and σ^2, respectively.

In addition, the required sample size computed with the classical formula [20, p. 46, equation 3.4] in order to obtain a power equal to $1 - \hat{\beta}_{\text{CR}}$ to detect the group effect γ has been computed as follows:

$$N_{0C} = \frac{(r+1) \times \hat{\sigma}^2 \times \left(z_{1-\alpha/2} + z_{1-\hat{\beta}_{\text{CR}}}\right)^2}{r\hat{\gamma}^2} \tag{11}$$

with

$$N_{1C} = rN_{0C}$$

and where the values of $\hat{\gamma}$ and $\hat{\sigma}^2$ are fixed to their expected planning values γ and σ^2, respectively.

In this case, the classical approach is used without distinction of the manifest or latent characteristic of the studied variables.

2.8. *Analysis and validation of the results*

We computed and compared the variance of the group effect (var_{CR}) obtained from the CRB with the mean variance of the estimations of the group effect obtained in the simulation study (var_S) computed as

$$\frac{1}{1000}\sum_{l=1}^{1000} \text{s.e.}^2(\hat{\gamma}_l) \tag{12}$$

where $\hat{\gamma}_l$ is the estimate of the γ parameter obtained in the lth replication of each case of the simulation study.

We compare the power defined by Equation (9) $(1 - \hat{\beta}_{\text{CR}})$with the power estimated in the simulation study $(1-\hat{\beta}_S)$ and with the power computed using the classical formula $(1-\hat{\beta}_C)$ defined in Equation (10).

We compare the sample size N with the required sample size N_C $(N_C = N_{0C} + N_{1C})$ classically computed using Equation (11).

2.9. *Practical computing of the Cramer–Rao bound*

In practice, the expected dataset obtained using the procedure that was previously described is used to determine the variance of γ (var_{CR}). A Rasch model including a group effect is fitted to this expected dataset with the parameters δ_j $(j = 1 \dots J)$ and σ^2 being fixed to their assumed values.

The STATA (StataCorp, College Station, TX, USA) [27] module -raschpower-, proposed by the authors, can be used to obtain the variance var_{CR} and the expected power $1 - \hat{\beta}_{\text{CR}}$. This module is stored on the Statistical Software Components of the Boston College Department of Economics (http://ideas.repec.org/s/boc/bocode.html) and can be directly downloaded from STATA with the command 'ssc install raschpower'.

The syntax of this module is as follows:

. raschpower [, n0(#) n1(#) gamma(#) variance(#) d(*string*)], where the #s are respectively the number of individuals in the group 0 (N_0, 100 by default), in the group 1 (N_1, 100 by default), the expected value of the group effect γ (0.5 by default), the expected value of the variance of the latent trait (σ^2, 1 by default), and the $string$ is the name of a vector containing the difficulty parameters of the items (δ_j) defined as the vector $(-1, -0.5, 0, 0.5, 1)'$ by default.

We realize the determination of the variance of the group effect using the -gllamm- module of the STATA 11 software [28] that allows fitting mixed logistic models. The -raschpower- module returns the variance (var_{CR}) and the standard error of the γ parameter based on the CRB, the expected value of the power $1 - \hat{\beta}_{\text{CR}}$ (Equation 9), the value of the power obtained using the classical formula $(1 - \hat{\beta}_C)$ (Equation 10), the required sample size computed with the classical formula (N_C for equal sample size in the two groups, that is, for $r = 1$) (Equation 11), and the ratio between N and N_C.

We present an example of an output of this module here.

```
. raschpower
Number of individuals in the first group:  100
Number of individuals in the second group: 100
Group effect: .5
Variance of the latent trait: 1
Number of items: 5
Difficulties parameters of the items:    -1  -.5    0   .5    1


--------------------------------------------------------------------------------
                                                    Estimation with the
                                           Cramer-Rao bound    classical formula
--------------------------------------------------------------------------------
Estimated value of the group effect                    0.52
Estimation of the s.e. of the group effect             0.20
Estimation of the variance of the group effect       0.0412
Estimation of the power                              0.6926               0.9424
Number of patients for a power of 69.26%            100/100        48.54/  48.54
Ratio of the number of patients                                2.06
--------------------------------------------------------------------------------
```

3. Results

Table I presents, for each scenario where $\gamma \neq 0$, the estimation of the variance of the group effect determined by the CRB (var_{CR}) and the mean value of this quantity obtained by simulations (var_S). These values are always very close to one another. We observe a decrease of the variance with the sample size and the number of items but a slight increase when the group effect increases.

Table II presents, for each values of J and $N_0 = N_1$, the estimated value of the type I error obtained with the simulation study. All the type I errors are close to the expected value of 5%, and none of them are significantly different from 5%. The type I error of the Wald test is well maintained at level α.

Table I. Variance of the group effect based on the Cramer–Rao bound (var_{CR}) and estimated by simulations (var_S) for different values of the group effect (γ), the sample sizes per group (N_0 and N_1 considered as equal), and the number of items J of the questionnaire (for a variance of the latent trait set at $\sigma^2 = 1$).

		γ			
		0.0	0.2	0.5	0.8
	$N_0 = N_1$	var_{CR} / var_S	var_{CR} / var_S	var_{CR} / var_S	var_{CR} / var_S
$J = 5$ items	50	0.0821 / 0.0822	0.0821 / 0.0822	0.0826 / 0.0825	0.0831 / 0.0832
	100	0.0410 / 0.0411	0.0411 / 0.0411	0.0412 / 0.0412	0.0416 / 0.0416
	200	0.0205 / 0.0205	0.0205 / 0.0205	0.0206 / 0.0206	0.0208 / 0.0208
	300	0.0137 / 0.0137	0.0137 / 0.0137	0.0137 / 0.0137	0.0138 / 0.0138
	500	0.0082 / 0.0082	0.0082 / 0.0082	0.0082 / 0.0082	0.0083 / 0.0083
$J = 10$ items	50	0.0642 / 0.0638	0.0642 / 0.0639	0.0643 / 0.0640	0.0647 / 0.0642
	100	0.0320 / 0.0319	0.0321 / 0.0319	0.0321 / 0.0320	0.0323 / 0.0321
	200	0.0160 / 0.0159	0.0160 / 0.0160	0.0161 / 0.0160	0.0161 / 0.0161
	300	0.0107 / 0.0106	0.0106 / 0.0106	0.0107 / 0.0107	0.0107 / 0.0107
	500	0.0064 / 0.0064	0.0064 / 0.0064	0.0064 / 0.0064	0.0064 / 0.0064

Table II. Type I error estimated by simulations for different values of the sample sizes per group (N_0 and N_1 considered as equal) and the number of items J of the questionnaire (for $\gamma = 0$ and a variance of the latent trait set at $\sigma^2 = 1$).

$N_0 = N_1$	$J = 5$ items	$J = 10$ items
50	0.047	0.046
100	0.038	0.061
200	0.058	0.041
300	0.047	0.048
500	0.054	0.053

Table III presents, for each scenario, the estimation of the power ($1 - \hat{\beta}_{\mathrm{CR}}$) of the Wald test based on the estimated value of the CRB, the value of the power obtained by simulation ($1-\hat{\beta}_S$), and the expected value of the power computed with the classical formula ($1 - \hat{\beta}_C$). We observe that $1 - \hat{\beta}_{\mathrm{CR}}$ and $1 - \hat{\beta}_S$ are close to one another, whatever the values of N, J, and γ; the difference ranges between -0.023 and 0.025 with a mean almost equal to 0. Concerning the power obtained by the classical formula, it is always higher than $1 - \hat{\beta}_{\mathrm{CR}}$ and $1 - \hat{\beta}_S$ (differences between $1 - \hat{\beta}_C$ and $1 - \hat{\beta}_{\mathrm{CR}}$ ranges between 0.000 and 0.288 with a mean at 0.099). The powers increase with N and γ. The powers $1 - \hat{\beta}_{\mathrm{CR}}$ and $1 - \hat{\beta}_S$ increase with J, but this is not the case for $1 - \hat{\beta}_C$ that remains constant whatever the value of J.

Table IV presents the required sample size per group ($N_{0C} = N_{1C} = N_C/2$) computed with the classical formula (Equation 11) to achieve the same power as the one obtained using the CRB ($1 - \hat{\beta}_{\mathrm{CR}}$) and the ratio between the total sample sizes used in the simulations and the ones obtained with the classical formula (N/N_C). We observe that the classical formula gives a smaller required sample size than the one used in the simulations to obtain the same specified value of power. The ratio between the two sample sizes decreases as J increases but seems to be stable for a given number of items with the same difficulty parameters, whatever the values of N. The value of the group effect (γ) seems to have a small impact on this ratio.

4. An example of sample size determination in a clinical context

We illustrate the results of this paper with an example coming from a pilot study whose data were used for sample size calculations for the planning of a future larger study [18]. The main objective of the

Table III. Powers computed from the Cramer–Rao bound ($1 - \hat{\beta}_{\mathrm{CR}}$) obtained by simulations ($1 - \hat{\beta}_S$) and expected with the classical formula ($1 - \hat{\beta}_C$) for the Wald test comparing the mean values of the latent trait in the two groups for different values of the group effect ($\gamma \neq 0$), sample sizes per group (N_0 and N_1 considered as equal), and number of items J of the questionnaire (for a variance of the latent trait set at $\sigma^2 = 1$ and $\alpha = 5\%$).

		γ		
		0.2	0.5	0.8
	$N_0 = N_1$	$1-\hat{\beta}_{\mathrm{CR}}/1-\hat{\beta}_S/1-\hat{\beta}_C$	$1-\hat{\beta}_{\mathrm{CR}}/1-\hat{\beta}_S/1-\hat{\beta}_C$	$1-\hat{\beta}_{\mathrm{CR}}/1-\hat{\beta}_S/1-\hat{\beta}_C$
$J = 5$ items	50	0.107/0.096/0.169	0.413/0.399/0.705	0.792/0.817/0.979
	100	0.167/0.169/0.293	0.693/0.675/0.942	0.975/0.977/1.000
	200	0.287/0.285/0.516	0.936/0.930/0.999	1.000/1.000/1.000
	300	0.401/0.409/0.688	0.989/0.990/1.000	1.000/1.000/1.000
	500	0.598/0.583/0.885	1.000/0.999/1.000	1.000/1.000/1.000
$J = 10$ items	50	0.124/0.117/0.169	0.505/0.511/0.705	0.882/0.872/0.979
	100	0.201/0.223/0.293	0.797/0.813/0.942	0.994/0.993/1.000
	200	0.353/0.330/0.516	0.977/0.976/0.999	1.000/1.000/1.000
	300	0.492/0.498/0.688	0.998/0.998/1.000	1.000/1.000/1.000
	500	0.706/0.721/0.885	1.000/1.000/1.000	1.000/1.000/1.000

Table IV. Power determined from the Cramer–Rao bound ($1-\hat{\beta}_{\text{CR}}$), sample sizes (per group) obtained with the classical formula ($N_{0C} = N_{1C} = N_C/2$) for a power ($1-\hat{\beta}_{\text{CR}}$) and ratio between the total sample size used in the simulations and the one computed with the classical formula (N/N_C), for different values of the group effect ($\gamma \neq 0$), sample sizes per group (N_0 and N_1 considered as equal), and number of items J of the questionnaire (for a variance of the latent trait set at $\sigma^2 = 1$ and $\alpha = 5\%$).

		γ		
		0.2	0.5	0.8
	$N_0 = N_1$	$1-\hat{\beta}_{\text{CR}}/\frac{N_C}{2}/\frac{N}{N_C}$	$1-\hat{\beta}_{\text{CR}}/\frac{N_C}{2}/\frac{N}{N_C}$	$1-\hat{\beta}_{\text{CR}}/\frac{N_C}{2}/\frac{N}{N_C}$
$J = 5$ items	50	0.107/ 24.36/2.05	0.413/ 24.22/2.06	0.792/ 24.06/2.08
	100	0.167/ 48.70/2.05	0.693/ 48.54/2.06	0.975/ 48.10/2.08
	200	0.287/ 97.36/2.05	0.936/ 97.05/2.06	1.000/ 96.34/2.08
	300	0.401/146.14/2.05	0.989/145.53/2.06	1.000/144.42/2.08
	500	0.598/243.52/2.05	1.000/242.54/2.06	1.000/240.71/2.08
$J = 10$ items	50	0.124/ 31.14/1.61	0.505/ 31.09/1.61	0.882/ 30.90/1.62
	100	0.201/ 62.33/1.60	0.797/ 62.31/1.60	0.994/ 62.00/1.61
	200	0.353/124.97/1.60	0.977/124.68/1.60	1.000/124.45/1.61
	300	0.492/187.91/1.60	0.998/187.44/1.60	1.000/186.50/1.61
	500	0.706/313.29/1.60	1.000/312.61/1.60	1.000/310.88/1.61

upcoming study is to compare the level of pain between two groups of patients having muscular dystrophies. The first group concerns patient with a Steinert's disease, and the second group on patients having another muscular dystrophy.

In the pilot study, the researchers recruited 52 patients with a Steinert's disease and 95 patients with another muscular dystrophy. They used the Nottingham Health Profile (NHP) questionnaire to evaluate the global QoL of the patients. In the upcoming study, we focus on the evaluation of pain. The pain dimension of the NHP is composed of eight binary items. In the pilot study, the researchers estimated the difficulty parameters at $(2.61, 2.94, 1.75, 0.46, -.11, 0.36, 1.28, 2.23)$. They estimated the variance of the latent trait 1.983^2 and the difference between the means in the two groups of patients at 0.649. The difference between the two means was not significant at 5% in this pilot study, but the p-value ($p = 0.08$) might suggest a possible lack of power for this study. So, it was interesting to determine a sample size to be able to significantly detect such a difference considered as clinically relevant.

The classical formula determines a sample size of 197 patients in each group for $\alpha = 5\%$ and $1-\beta = 90\%$. The -raschpower- STATA module has been used with these values.

The -raschpower- STATA module predicts a power of only 80% to detect a difference on the means of the latent trait between the two groups with this sample size using a Rasch model. The classical formula predicts only 148 patients per groups to obtain such a power. The ratio between the two sample sizes is estimated at 1.34. The results obtained with the simulation study designed in this paper show that this ratio is constant for a given vector of difficulty parameters, whatever the sample size. Consequently, it is possible to use this ratio to predict an accurate sample size from the sample size obtained with the classical formula: the correct sample size is approximated at $197 \times 1.34 = 264$ patients per group. The -raschpower- STATA module is run with this new value.

The power estimated with this new sample size is estimated at 90.22%, which is very close to the expected value of 90%.

To confirm these results, we simulate 1000 datasets with eight items whose difficulties are equal to the ones used with -raschpower-. We used a normal distribution of variance 1.983^2 with a mean equal to $-0.649/2$ for the first group and $0.649/2$ for the second group to simulate the latent trait. We conducted this short simulation study with 197 and 264 individuals for the first and second groups, respectively. We fitted each simulated dataset using a Rasch model considering the difficulty parameters and the value of the variance of the latent trait as known and estimating the group parameter. With 197 individuals per group, the power obtained on the Wald test on the group effect is equal to 80.5%, and with 264 individuals per group, this power is equal to 89.7%. These results confirm the possibility to use such a process to determine a sample size in the context of the Rasch model.

5. Discussion

We proposed a theoretical approach for IRT-based sample size determination for two-group cross-sectional comparisons. It takes into account item parameters, the minimum clinically relevant difference expressed as the difference of the means of the latent traits as well as its variance whose derivation is approximated using Fisher's information and CRB property. We computed the power and the corresponding sample sizes obtained using this strategy and the usual one based on the classical formula for normally distributed endpoints and compared for a variety of situations often encountered in clinical and epidemiological research.

The estimate of the variance of the difference between the means of the latent traits in the two groups obtained using the proposed methodology is supposed to achieve equality on Cramer–Rao's inequality if it is efficient and hence corresponds to the minimum variance unbiased estimator. MML, which is used for estimation, is known to provide asymptotically efficient estimators [26]; thus, it might be hypothesized that equality of the CRB is attained at least asymptotically.

Comparisons of sample sizes and corresponding power computed using either the classical sample size formula developed for manifest normally distributed endpoints or the proposed IRT-based strategy illustrate the fact that the classical formula is inadequate for IRT models. Indeed, performing such sample size calculations leads to an underestimation of the size of the study if IRT models are intended to be used for analysis of PRO data and hence a substantial loss in power. In particular, a strong impact of the number of items in the questionnaire on power has been observed: the power increases with the length of the studied questionnaire (J). This constitutes an important issue because the classical formula does not take into account the size of the questionnaire (J). This point has already been stressed [29, 30] as well as the fact that, as the number of items increases, the power obtained using IRT modeling seems to progressively attain the power associated with the classical formula [18]. This could represent a theoretical situation where the precision of the measure of the latent trait is very good, and therefore might correspond to the case where the latent variable could have characteristics similar to those of an observed variable.

Comparison of the required sample size computed using both formulas for a given value of power shows that the ratio between the sample size using CRB and the one obtained by the classical formula was higher than 1. Moreover, this ratio seems to depend on the number of items (and on their difficulty parameters—results not shown) and also but more slightly on the value of the group effect γ. However, the study size and the value of the desired power do not seem to have an impact on this ratio. A further topic will be to determine whether a correction coefficient depending on these parameters and this ratio would allow a simple modification of the classical formula that could make it suitable for latent variables. The next step could be to determine an analytic formula for this ratio. Meanwhile, the -raschpower-module can provide a reliable numerical approximation because this ratio seems to be very stable for a given vector of items difficulties and value of γ, whatever the sample size. The illustrative example proposed in the paper shows a nice and easy way to determine a suitable sample size for latent variables.

The main drawbacks of the proposed approach are its complexity compared with the classical formula and the fact that the practical implementation of this approach is based on an estimation of the group effect γ, even if during the planning of a study, all the parameters are assumed to be known and are fixed. We also explored the impact of using an estimate of γ instead of a fixed value on the estimated bias of the power in the 30 tested scenarios. On the one hand, we observed that the absolute difference between $1 - \hat{\beta}_{\mathrm{CR}}$ and $1 - \hat{\beta}_S$ was small (it ranged between 0.000 and 0.025 with a mean at 0.007 and a median at 0.004). These differences generally are not relevant in practice. On the other hand, a linear model explaining the absolute differences $|(1 - \hat{\beta}_{\mathrm{CR}}) - (1 - \hat{\beta}_S)|$ as a function of $|\hat{\gamma} - \gamma|$ adjusted or not on the sample size (N), the number of items (J), and the real value of the group effect (γ) was fitted and showed a nonsignificant effect of $|\hat{\gamma} - \gamma|$ ($p = 0.22$ without adjustment and $p = 0.21$ after adjustment). It is therefore expected that using a fixed value for γ or estimating it should have a very slight impact on power $1 - \hat{\beta}_{CR}$ and on the associated sample size. However, it could be important to improve the -raschtest- STATA module to be able to estimate the power without such an estimation of the γ parameter.

The knowledge of the difficulty parameters of the items can be considered as another drawback. Nevertheless, we can stress that nowadays (i) many PRO instruments are also validated using IRT models; (ii) items banks are being constituted [14]; and/or (iii) previous data from a pilot study might also be used as we did in our example. Hence, estimation of item parameters coming from validated PRO can be obtained and used for IRT-based sample size calculation. Moreover, Sébille *et al.* [18] realized a simulation study in order to estimate the power using IRT with known or unknown difficulty parameters. They showed that the power is slightly affected by a relatively poor precision of the parameter (±1 compared with the real values of the parameters). Hence, precise knowledge of the difficulty parameters seems not to be essential, and we might expect to be able to estimate the power of a study with a good enough precision.

An extension of this work concerns the fact that in the present approach, the difficulty parameters and the variance of the latent trait are considered as known and are not estimated jointly with the γ parameter. In practice, it could be preferable to fix the values of the difficulty parameters to be comparable between different studies. However, concerning the variance of the latent trait (σ^2), this parameter is rarely known with good precision, and it is systematically (re)estimated. The approach will be developed to give the possibility to estimate jointly γ and σ^2 (and eventually the difficulty parameters as well). This development is complex because correlations between the different estimators will appear in the analytical development, but it could be very useful in practice.

Another extension of this work concerns the development of this approach to other IRT models. Its implementation for other dichotomous models like the Birnbaum model [31] or the one parameter logistic model (OPLM) [21] seems easy. These two models allow weighting the items for the estimation of the latent trait (in the Rasch model, the contribution of each item to the estimation of the latent trait is the same). The difference between the Birnbaum model and the OPLM is that these weights can be estimated (Birnbaum model) or fixed by the user (OPLM). In the case of the OPLM, these weights are known a priori, and so it is easy to apply the procedure described in this paper. In the Birbaum model, the user must have an idea of these weights that have to be estimated. For planning purposes, these weights are fixed to assumed values, as for the difficulty parameters of the Rasch model.

The extension to polytomous models like the partial credit model [24] or the rating scale model [23] can reveal some other difficulties notably concerning the determination of the expected dataset because for such models, the number of response patterns could be very important, even with a small number of items. Moreover, with the partial credit model, there is one difficulty parameter per category of response to each item, and so the number of parameters can be large, and it could be difficult to have a good idea of an appropriate value for each of them.

Extending the proposed approach to other designs often used for PRO data such as longitudinal studies would also be worthwhile. Longitudinal IRT models [30] could be used for this purpose to provide valid sample size methodology for testing a time effect in the case of a one-sample design (single group) or a time effect, a group effect and possible interaction between them in the case of a two-sample design (two independent groups) or more.

Some other aspects related to the proposed approach could also be investigated more thoroughly. In particular, it could be interesting to investigate potential effects of some parameters on the expected power based on the CRB, notably the impact of the values of the difficulty parameters of the items in terms of mean location and dispersion of these parameters. Moreover, despite great efforts, clinicians and statisticians often face a high degree of uncertainty on some parameters when designing a study on PRO endpoints to justify sample size. Therefore, it could be valuable to investigate the influence of the precision of the values of the main parameters (difficulty parameters, group effect, and variance of the latent trait) on the obtained power. Indeed, the impact of fixing incorrect values for these parameters is an important topic for practical use of this approach, where these values are not always known with very good precision.

APPENDIX A.

Let $l = \log(L)$, we wish to obtain the second derivative with respect to γ of the log-likelihood l, we shall in the following use l instead of $l(\sigma^2, \delta_1, ..., \delta_J, \gamma|x)$. Recall that because we are at the planning phase of a study, γ, σ^2, and $\delta_1, ..., \delta_J$ are fixed parameters set to some hypothesized values.

We can write the first derivative with respect to γ of the log-likelihood as

$$\frac{\partial l}{\partial \gamma} = \frac{\partial}{\partial \gamma} \log\left(\prod_{g=0}^{1} \prod_{n=1}^{N_g} \int \prod_{j=1}^{J} \frac{\exp(x_{nj}(\theta + (-1)^{1-g} N_{1-g}\gamma/N - \delta_j))}{1 + \exp(\theta + (-1)^{1-g} N_{1-g}\gamma/N - \delta_j)} G(\theta|\sigma^2)\mathrm{d}\theta \right) \qquad (13)$$

Let for $n=1,\dots,N$ and $g=0,1$:

$$f_{gn}=\int\prod_{j=1}^{J}\frac{\exp(x_{nj}(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j))}{1+\exp(\theta+(-1)^{1-g}N_{1-g}\gamma/N-\delta_j)}G(\theta|\sigma^2)\mathrm{d}\theta \tag{14}$$

Hence,

$$\frac{\partial l}{\partial\gamma}=\sum_{g=0}^{1}\sum_{n=1}^{N_g}\frac{\partial}{\partial\gamma}\log(f_{gn}) \tag{15}$$

We can write the second derivative with respect to γ of the log-likelihood as

$$\frac{\partial^2 l}{\partial\gamma^2}=\sum_{g=0}^{1}\sum_{n=1}^{N_g}\left(\frac{\partial^2 f_{gn}}{\partial\gamma^2}\times f_{gn}-\left(\frac{\partial f_{gn}}{\partial\gamma}\right)^2\right)\times\frac{1}{(f_{gn})^2} \tag{16}$$

If we denote f_n and f_{nj} the following functions ($n=1,\dots,N$ and $j=1,\dots,J$:

$$f_n=\int\prod_{j=1}^{J}\frac{\exp(x_{nj}(\theta+\gamma-\delta_j))}{1+\exp(\theta+\gamma-\delta_j)}G(\theta|\sigma^2)\mathrm{d}\theta) \tag{17}$$

where

$$f_{nj}=\frac{\exp(x_{nj}(\theta+\gamma-\delta_j))}{1+\exp(\theta+\gamma-\delta_j)} \tag{18}$$

We can write the first and second derivatives of f_{gn} ($g=0,1$) with respect to γ as a function of f_n ($n=1,\dots,N$):

$$\frac{\partial f_{gn}}{\partial\gamma}(\gamma)=(-1)^{1-g}N_{1-g}/N\times\frac{\partial f_n}{\partial\gamma}\left((-1)^{1-g}N_{1-g}\gamma/N\right) \tag{19}$$

$$\frac{\partial^2 f_{gn}}{\partial\gamma^2}(\gamma)=\left(N_{1-g}/N\right)^2\times\frac{\partial^2 f_n}{\partial\gamma^2}\left((-1)^{1-g}N_{1-g}\gamma/N\right) \tag{20}$$

where

$$\frac{\partial f_n}{\partial\gamma}=\int\left(\sum_{j=1}^{J}\frac{\partial f_{nj}}{\partial\gamma}\prod_{i\neq j,i=1}^{J}f_{ni}\right)G(\theta|\sigma^2)\mathrm{d}\theta. \tag{21}$$

$$\frac{\partial f_{nj}}{\partial\gamma}=\frac{x_{nj}\exp(x_{nj}(\theta+\gamma-\delta_j))+(x_{nj}-1)\exp((x_{nj}+1)(\theta+\gamma-\delta_j))}{(1+\exp(\theta+\gamma-\delta_j))^2} \tag{22}$$

and

$$\frac{\partial^2 f_n}{\partial\gamma^2}=\int\sum_{j=1}^{J}\left(\frac{\partial^2 f_{nj}}{\partial\gamma^2}\prod_{i=1,i\neq j}^{J}f_{ni}+\frac{\partial f_{nj}}{\partial\gamma}\sum_{k=1,k\neq j}^{J}\frac{\partial f_{nk}}{\partial\gamma}\prod_{l=1,l\neq k,l\neq j}^{J}f_{nl}\right)G(\theta|\sigma^2)\mathrm{d}\theta \tag{23}$$

with

$$\frac{\partial^2 f_{nj}}{\partial\gamma^2}=\frac{\left(x_{nj}^2\exp(x_{nj}(\theta+\gamma-\delta_j))+(x_{nj}^2-1)\exp((x_{nj}+1)(\theta+\gamma-\delta_j))\right)\times(1+\exp(\theta+\gamma-\delta_j))^2}{(1+\exp(\theta+\gamma-\delta_j))^4} \tag{24}$$

$$-\frac{2\exp(\theta+\gamma-\delta_j)(1+\exp(\theta+\gamma-\delta_j))[x_{nj}\exp(x_{nj}(\theta+\gamma-\delta_j))+(x_{nj}-1)\exp((x_{nj}+1)(\theta+\gamma-\delta_j))]}{(1+\exp(\theta+\gamma-\delta_j))^4} \tag{25}$$

Acknowledgement

The French National Research Agency, under reference N 2010 PRSP 008 01, supported this study.

References

1. Joly F, Vardy J, Pintilie M, Tannock IF. Quality of life and/or symptom control in randomized clinical rials for patients with advanced cancer. *Annals of Oncology* 2007; **18**:1935–1942. DOI: 10.1093/annonc/mdm121.
2. Bottomley A, Jones D, Claassens L. Patient-reported outcomes: assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency. *European Journal of Cancer* 2009; **45**:347–353. DOI: 10.1016/j.ejca.2008.09.032.
3. Walters SJ, Campbell MJ, Lall R. Design and analysis of trials with Quality of life as an outcome: a practical guide. *Journal of Biopharmaceutical Statistics* 2001; **11**:155–176. DOI: 10.1081/BIP-100107655.
4. Julious SA, George S, Machin D, Stephens RJ. Sample sizes for randomized trials measuring quality of life in cancer patients. *Quality of Life Research* 1997; **6**:109–117. DOI: 10.1054/bjoc.2000.1383.
5. Gotay CC, Lipscomb J, Snyder CF. Reflections on findings of the cancer outcomes measurement working group: moving to the next phase. *Journal of the National Cancer Institute* 2005; **97**:1568–1574. DOI: 10.1093/jnci/dji337.
6. Cella D, Beaumont JL, Webster KA, Lai JS, Elting L. Measuring the concerns of cancer patients with low platelets counts: the Functional Assessment of Cancer Therapy - Thrombocytopenia (FACT-Th) questionnaire. *Support Care Cancer* 2006; **14**:1220–1231. DOI: 10.1007/s00520-006-0102-1.
7. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, Brdart A, Fayers P, Jordhoy M, Sprangers M, Watson M, Young T. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. *Quality of Life Research* 2004; **13**:1683–1697. DOI: 10.1007/s11136-004-7866-x.
8. Garcia SF, Cella D, Clauser SB, Flynn KE, Lad T, Lai JS, Reeve BB, Smith AW, Stone AA, Weinfurt K. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. *Journal of Clinical Oncology* 2007; **25**:5106–5112. DOI: 10.1200/JCO.2007.12.2341.
9. Siu AM, Lai CY, Chiu AS, Yip CC. Development and validation of a fine-motor assessment tool for use with young children in a Chinese population. *Research in Developmental Disabilities* 2011; **32**:107–114. DOI: 10.1016/S1569-1861(10)70007-3.
10. Brady CJ, Keay L, Villanti A, Ali FS, Gandhi M, Massof RW, Friedman DS. Validation of a visual function and quality of life instrument in an urban Indian population with uncorrected refractive error using Rasch analysis. *Ophthalmic Epidemiology* 2010; **17**:282–291. DOI: 10.3109/09286586.2010.511756.
11. Brown T, Unsworth C, Lyons C. An evaluation of the construct validity of the Developmental Test of Visual-Motor Integration using the Rasch Measurement Model. *Australian Occupational Therapy Journal* 2009; **56**:393–402. DOI: 10.1111/j.1440-1630.2009.00811.x.
12. Hayas CA, Quintana JM, Padierna JA, Bilbao A, Muoz P. Use of Rasch methodology to develop a short version of the Health Related Quality of life for Eating Disorders questionnaire: a prospective study. *Health and Quality of Life Outcomes* 2010; **8**:29. DOI: 10.1186/1477-7525-8-29.
13. Hart DL, Deutscher D, Wernekeyy MW, Holder J, Wang YC. Implementing computerized adaptive tests in routine clinical practice: experience implementing CATs. *Journal of Applied Measurements* 2010; **11**:288–303.
14. Gershon RC, Rothrock N, Hanrahan R, Bass M, Cella D. The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *Journal of Applied Measurements* 2010; **11**:304–314.
15. Kisala PA, Tulsky DS. Opportunities for CAT applications in medical rehabilitation: development of targeted item banks. *Journal of Applied Measurements* 2010; **11**:315–330.
16. Petersen MA, Groenvold M, Aaronsony NK, Chie WC, Conroy T, Costantini A, Fayers P, Helbostady J, Holznery B, Kaasa S, Singer S, Velikova G, Youngy T. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. *Quality of Life Research* 2011; **20**:479–490. DOI: 10.1016/j.ejca.2010.02.011.
17. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, Gregg PJ. A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Quality of Life Research* 2004; **13**:331–338. DOI: 10.1023/B:QURE.0000018489.25151.e1.
18. Sébille V, Hardouin JB, Le Néel T, Kubis G, Boyer F, Guillemin F, Falissard B. 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 simulation study. *BMC Medical Research Methodology* 2010; **10**:24. DOI: 10.1186/1471-2288-10-24.
19. Draxler C. Sample size determination for Rasch model tests. *Psychometrika* 2010; **75**:708–724. DOI: 10.1007/s11336-010-9182-4.
20. Julious SA. *Sample Sizes for Clinical Trials*. CRC Press: Boca Raton, 2010.
21. Fishery GH, Molenaar IW. *Rasch Models, Foundations, Recent Developments, and Applications*. Springer-Verlag: New-York, 1995.
22. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press: Chicago, 1980.
23. Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43**:561–573. DOI: 10.1007/BF02293814.
24. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; **47**:149–174. DOI: 10.1007/BF02296272.
25. 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. DOI: 10.2307/1390625.

26. Thissen D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* 1982; **47**:175–186. DOI: 10.1007/BF02296273.
27. StataCorp. *Stata: Release 11. Statistical Software*. Stata Press: College Station, TX, 2009.
28. Rabe-Hesketh S, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata*, 2nd ed. Stata Press: College Station, TX, 2008.
29. Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. *Controlled Clinical Trials* 2003; **24**:390–410. DOI: 10.1016/S0197-2456(03)00061-8.
30. Glas CAW, Geerlings H, van de Laar MAFJ, Taal E. Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials* 2009; **30**:158–170. DOI: 10.1016/j.cct.2008.12.003.
31. Birnbaum A. Some latent trait models and their use in inferring an examinees ability. In *Statistical Theories of Mental Test Scores*, Lord FM, Novick MR (eds). Addison-Wesley: Reading, MA, 1968.