Statistics
in Medicine

Research Article

Received 7 July 2013, Accepted 23 February 2015 Published online 18 March 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6478

Power and sample size determination for group comparison of patient-reported outcomes using polytomous Rasch models

Jean-Benoit Hardouin,[a,b*†] Myriam Blanchin,[a] Mohand-Larbi Feddag,[a] Tanguy Le Néel,[a] Bastien Perrot[a,b] and Véronique Sébille[a,b]

The analysis of patient-reported outcomes or other psychological traits can be realized using the Rasch measurement model. When the objective of a study is to compare groups of individuals, it is important, before the study, to define a sample size such that the group comparison test will attain a given power. The Raschpower procedure (RP) allows doing so with dichotomous items. The RP is extended to polytomous items. Several computational issues were identified, and adaptations have been proposed. The performance of this new version of RP is assessed using simulations. This adaptation of RP allows obtaining a good estimate of the expected power of a test to compare groups of patients in a large number of practical situations. A Stata module, as well as its implementation online, is proposed to perform the RP. Two versions of the RP for polytomous items are proposed (deterministic and stochastic versions). These two versions produce similar results in all of the tested cases. We recommend the use of the deterministic version, when the measure is obtained using small questionnaires or items with a few number of response categories, and the stochastic version elsewhere, so as to optimize computing time. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Rasch model; Raschpower; sample size; partial credit model

1. Introduction

The reporting of a patient's own perception of his/her health condition has gained much attention. Hence, endpoints such as quality of life (QoL) and other perceived health measures (pain, fatigue, and etc.) are increasingly used as important health outcomes in clinical trials and epidemiological studies, and are considered highly valued endpoints of medical care in different areas, for example, rheumatology, oncology, cardiology, and surgery [1–7]. These types of endpoints are usually referred to as latent variables or latent traits. They cannot be directly observed nor measured as other clinical or biological data, and they are often collected through self-assessment questionnaires including either binary or polytomous items and are termed patient-reported outcomes (PRO). The patient's responses to the items are often combined to provide scores that are subsequently used for analysis with the so-called classical test theory (CTT) analysis. An additional possibility that has gained a great deal of interest during the past years [8,9] is to work directly on the item responses by fitting models based on the item response theory (IRT), in particular, the Rasch model [10] for binary items or the partial credit model (PCM) for polytomous items [11]. These models enable to model relationships between the observed variables representing the answers to the items and latent variables of interest (QoL, anxiety, and etc.). Many PRO instruments are

[a]*EA4275 - SPHERE "Biostatistics, Pharmacoepidemiology and Human Sciences Research", Faculties of Medicine and Pharmaceutical Sciences, University of Nantes - PRES UNAM 1 rue Gaston Veil, 44035 Nantes, France*
[b]*Team of Methodology and Biostatistics, University Hospital of Nantes - Clinical Research, Unit 1 place Alexis-Ricordeau, 44093 Nantes, France*
[*]*Correspondence to: Jean-Benoit Hardouin, EA4275 "Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences", Faculties of Medicine and Pharmaceutical Sciences, University of Nantes - PRES UNAM, 1 rue Gaston Veil, 44035 Nantes, France.*
[†]*E-mail: jean-benoit.hardouin@univ-nantes.fr*

currently found to be well adapted to IRT modeling because such models are used more and more for the construction, validation, and reduction of questionnaires [12, 13].

Sample size determination is an issue of primary importance during the planning of a study. Indeed, studies of inadequate size may lead to erroneous and uninformative conclusions, and may expose patients to inappropriate medical strategies. Sample size calculations have been widely developed for many types of endpoints, and specific methods exist for sample size determination for a number of outcomes including quantitative, categorical, and censored data [14–16]. All of the methods used for sample size planning and for statistical analysis are based on similar grounds. For example, the well-known sample size formula for the comparison of normally distributed endpoints between two groups of patients is based on the *t*-test [14]. However, for the comparison of PRO data in cross-sectional studies, it has been recently emphasized that this sample size formula was inadequate if a Rasch model was intended to be used for analysis. Indeed, using this formula leads to an underestimated sample size and hence poor power [17]. Some methodological developments have therefore been proposed for sample size calculations for the comparison of PRO data in two groups of patients using the Rasch model [18]. In this approach, named Raschpower, the expected item parameters of the PRO, the difference in the latent variables means, and its variance, for which the derivation was approximated using Cramer–Rao (CR) bound, were all taken into account. The approximation of this latter variance required the determination of the expected patient responses to the binary items, that is, the 2^J (where J is the number of items) possible response patterns and their associated probabilities of occurrence computed using the Rasch model. The expected dataset that was obtained was then used to determine the CR bound, and this approach provided adequate power and sample size calculations for PRO including binary responses [18].

However, many PROs in health science for the most part include polytomous rather than dichotomous items, and this paper aims at providing the necessary developments to adapt the previous methodology to such data. One of the main issues for this extension is to take into account the huge number of possible response patterns for which the computations are required for sample size and power determinations. Four different strategies are proposed to handle this problem and are compared using simulations, regarding power and sample size.

2. Methods

2.1. The latent regression-mixed partial credit model

The objective of this work is to provide power calculations for the comparison of PRO measured by questionnaires composed of polytomous items between two groups of patients when the data are intended to be subsequently analyzed using ordinal Rasch models such as the PCM.

In Rasch models, the latent trait can be considered a set of fixed parameters (with one parameter per individual reflecting a measure of the underlying concept for each individual) or as a random variable. In the latter case, we consider the sample to be representative of a more general population, and an assumption is made for the distribution of the random variable. The parameters of this distribution are then estimated. In general, the latent variable is assumed to be normally distributed, and we estimate the mean and the variance of this distribution. In this approach, the analysis is performed at the population level but not at an individual level, and covariates can be added to characterize the individuals in the population. For example, in clinical research, the treatment received by the patients allows for the ability to distinguish between groups of patients. In each group, the mean and variance of the latent trait can be estimated, and the groups can be compared using the means.

A well-known model of the Rasch family for polytomous items is the PCM [11]. This model is quite similar to the adjacent category logit model used to model ordinal data. The main difference is that the latent trait is introduced as a random variable among the covariates. In the PCM, the probability of answering the hth response category ($h = 0, \dots, K$) of the jth item X_j ($j = 1, \dots, J$) with $K + 1$ response categories is modeled as

$$P\left(X_j = h|\theta;\delta_j\right) = \frac{\exp\left(h\theta - \sum_{l=1}^{h}\delta_{jl}\right)}{1 + \sum_{r=1}^{K}\exp\left(r\theta - \sum_{l=1}^{r}\delta_{jl}\right)}, \tag{1}$$

with θ as a specific value of the latent trait distributed according to a normal distribution with means μ_0 and μ_1, and variances σ_0^2 and σ_1^2 (for the two groups indexed by 0 and 1, respectively). $\delta_j = (\delta_{j1} \dots \delta_{jr} \dots \delta_{jm_j})$

is a vector of parameters defining the difficulties of the jth item having $m_j + 1$ answer categories. This principle of latent regression can be adapted to others models of the IRT [19–21].

The main criterion is the group parameter defined as the difference between the two means denoted as $\gamma = \mu_1 - \mu_0$. It is typical to compare the group parameter with 0 using a Wald test. Because a constraint must be defined to identify the model, let $\mu = 0$ with μ the mean between μ_0 and μ_1; each of them weighted by the sample sizes N_0 and N_1 in each group. Consequently,

$$\begin{cases} N_0\mu_0 + N_1\mu_1 = 0 \\ \gamma = \mu_1 - \mu_0 \end{cases} \Leftrightarrow \begin{cases} \mu_0 = -\frac{N_1\gamma}{N_0+N_1} \\ \mu_1 = \frac{N_0\gamma}{N_0+N_1}. \end{cases}$$

In the particular case in which the two groups are of equal size ($N_0 = N_1$), we obtain $\mu_0 = -\gamma/2$ and $\mu_1 = \gamma/2$. The hypotheses tested by the Wald test can be written as $H_0 : \gamma = 0$ and $H_1 : \gamma \neq 0$.

The parameters of the model can be estimated using marginal maximum likelihood by maximizing the quantity

$$l_M(\delta, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2|x) = \prod_{n=1}^{N_0+N_1} \int_{-\infty}^{+\infty} \prod_{j=1}^{J} P(X_{nj} = x_{nj}|\theta_n; \delta_j) G(\theta|\mu_0, \mu_1, \sigma_0^2, \sigma_1^2) d\theta, \tag{2}$$

with G(.) as the density function of the normal distribution and δ as a vector of all item parameters δ_{jl}, $j = 1, \ldots, J$ and $l = 1, \ldots, K$, and x the set of responses of the $N_0 + N_1$ individuals to the J items. Let Γ be the estimator of γ, which is the difference between the estimates of μ_0 and μ_1 [22]. The statistic of the Wald test is defined by $\Gamma/\sqrt{Var(\Gamma)}$ and follows, under the H_0 hypothesis, a standardized normal distribution for a large sample size.

2.2. *Estimation of the power using Cramer–Rao bound: the Raschpower procedure*

During the planning step, a sample size is determined to detect a relevant clinical difference between groups, γ, with a power $1 - \beta$ at a given α level. The assumptions needed to compute this sample size include the pre-specification of the values of the item parameters, δ, and the variances of the latent trait σ_0^2 and σ_1^2, which are *a priori* set to the so-called planned values (expected values of these parameters can be obtained from previous studies for example). The standard error of γ is also needed and requires additional assumptions regarding the pattern of patient's responses to the items. The main difficulty is related to the approximation of the standard error of the γ estimate. Indeed, this estimation requires a dataset of the answers of the individuals to the items, but during the planning step, this dataset is, by definition, unavailable. An expected dataset is therefore created to avoid this difficulty. By analogy with [18], it is approximated at the planning step using a theoretical dataset created using the expected values of the parameters (δ, σ_0^2, σ_1^2 and γ) as well as the model that will be used to analyze the data, the PCM.

To do so, all of the possible response patterns are created, and for each of them, two probabilities are computed, $\pi_0(\mathbf{x})$ and $\pi_1(\mathbf{x})$, corresponding to the probability of observing the response pattern $\mathbf{x}$ for an individual of the first group and the second group, respectively. These two probabilities are approximated using the PCM. Using the local independence property, these probabilities can be written as

$$\pi_g(\mathbf{x}) = \int_{-\infty}^{+\infty} \prod_{j=1}^{J} P(X_j = x_j|\theta; \boldsymbol{\delta}_\mathbf{j}) G(\theta|\mu_g, \sigma_g^2) d\theta \quad \forall g = \{0, 1\}, \tag{3}$$

with $G(\theta|\mu_g, \sigma_g^2)$ as the normal density function with mean μ_g and variance σ_g^2. Each of these values (for each group $g = \{0, 1\}$ and for each possible K^J response patterns) can be computed using Gauss–Hermite quadratures (GM method).

For each response pattern, the expected number of associated individuals per group is computed as $n_{g,\mathbf{x}} = \mathrm{floor}(n_g \times \pi_g(\mathbf{x}))$, where n_g is the number of individuals of group g, and $\mathrm{floor}(x)$ is a function such that $\mathrm{floor}(x) = n$ if $n \leqslant x < n + 1$ with n an integer. As each number is rounded to the closest smaller integer, the assigned number of individuals among all of the possible response patterns is less than the expected number in each group (n_g, $g = \{0, 1\}$). To allocate the n_g individuals in each group, the results are incremented by 1 for the patterns having, after this first assignment, the greatest values $n_g \times \pi_g(\mathbf{x}) - n_{g,\mathbf{x}}$ until obtaining $\sum_\mathbf{x} n_{g,\mathbf{x}} = n_g$ for each group ($g = \{0, 1\}$).

This procedure (named GH) is similar to the one proposed in [18] for dichotomous items, but it is faced with the problem of the huge number of possible response patterns that can be obtained with polytomous

items. Indeed, if the scale is composed of J items with K response categories, the number of possible responses is K^J; a number that can become very large. For example, for 10 items with seven responses categories, the number of response patterns, 7^{10}, is close to 300 million. This creates computational issues because it is not easy to handle several million data; furthermore, the computing time can become very large to approximate all of the values $\pi_g(\mathbf{x})$ by GH method.

Three adaptations of the GH method are proposed to reduce computing time:

- Mean method: The quantity $\pi_g(\mathbf{x})$ is approximated by using only the mean of θ (and not the entire distribution), such as

$$\hat{\pi}_g(\mathbf{x}) = \prod_{j=1}^{J} P(X_j = x_j | \theta_g = \mu_g, \boldsymbol{\delta}_\mathbf{j}) \; \forall g = \{0, 1\}. \tag{4}$$

- Mean + GH method: All of the $\pi_g(\mathbf{x})$ are approximated by the mean method, and then these quantities are approximated by the GH method only for the P_g response patterns having the greatest values for $\hat{\pi}_g(\mathbf{x})$. We propose to use $P_g = 2 \times n_g$.
- Population + GH method: A large dataset with T individuals is simulated using a PCM, and then the $\pi_g(\mathbf{x})$ are approximated by GH method only for the most frequent P_g response patterns in this simulated dataset. We propose to use $T = 1,000,000$ and $P_g = 2 \times n_g$.

When each studied response pattern $\mathbf{x}$ is associated with a frequency $n_{g,\mathbf{x}}$, an expected dataset is created and analyzed by a latent regression-mixed partial credit model [23] with a random effect for the latent trait. Variances of the latent trait and difficulty parameters of the items are set to their planned values. The difference γ between the two mean values of the latent trait in each group ($\gamma = \mu_1 - \mu_0$) is estimated, and its variance ($var(\hat{\gamma})$) is approximated using the CR Bound. Finally, the power of the test based on this estimation of the CR bound is approximated at

$$1 - \beta_{\text{CR}} \approx 1 - \Phi\left(z_{1-\alpha/2} - \frac{|\gamma|}{\sqrt{var(\hat{\gamma})}}\right), \tag{5}$$

with $\Phi(.)$ as the cumulative standard normal distribution function.

2.3. Estimation of the power using simulated datasets

The estimation of the power based on the CR bound ($1 - \beta_{\text{CR}}$) is compared with the estimation obtained using simulated datasets ($1 - \beta_S$). A large number of datasets were created using the expected values of the parameters defined in the planning step as simulation parameters. The latent trait was then simulated using draws from a normal distribution, with the planned values for the mean μ_g, variance σ_g^2 for an individual of the group g and with N_g individuals for group g, $g = \{0, 1\}$. For each individual, it was possible to obtain the probabilities of having a given response for each item using the item response functions of the PCM (equation 1).

One thousand replications of each of the 90 cases (defined for given values for $J = \{5, 10\}$, $N_0 = N_1 = \{50, 100, 200, 300, 500\}$, $K = \{3, 5, 7\}$, $\gamma = \{0.2, 0.5, 0.8\}$, and $\sigma_1^2 = \sigma_2^2 = 1$) have been performed. Note that the Raschpower procedure runs with an odd or an even number of response categories K, but for the simulations, to explore the range of values between 3 and 7, only odd numbers have been used. Difficulty parameters are drawn in the $[-2; 2]$ interval with different values according to the number of items (J) and the number of response categories (K), with $\delta_{j1} < \delta_{j2} < \cdots < \delta_{jK}, \forall j = 1, \ldots, J$. For each simulated dataset, the γ parameter is estimated with its variance; a Wald test of the null hypothesis $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ is realized. The power estimated from the simulation study ($1 - \beta_S$) is defined for each case as the rate of significant Wald tests under H_1.

Because, in practice, the variance of the latent trait is rarely fixed when analyzing data using a Rasch model and because the difficulty parameters are sometimes unknown; three cases were considered for simulations:

- In the first case (S1), the variance of the latent trait and the difficulty parameters were considered known to be comparable with the assumptions of the Raschpower procedure;
- In the second case (S2), difficulty parameters were considered known, but the variance of the latent trait was estimated jointly with the group effect;
- In the third case (S3), difficulty parameters and variance of the latent trait were estimated jointly with the group effect.

2.4. The Raschpower Stata and online modules

The Raschpower procedure based on the CR bound can be performed using the Raschpower Stata module, which can be downloaded from http://raschpower.anaqol.org. This module is also implemented online on the website PRO-online (http://pro-online.univ-nantes.fr/). Consequently, it is possible to use the procedure independently from the usual statistical software of the user.

The general syntax of this Stata module is as follows:

raschpower [, Difficulties(*matrix*) n0(*integer*) n1(*integer*) Gamma(*real*) Var(*string*) Method(*string*) NBPatterns(*integer*) EXPectedpower(*real*)].

A matrix of the difficulty parameters must be previously defined. This matrix will have as many columns as items and as many rows as response categories to each item and will contain each of the difficulty parameters. Such a matrix must be presented as follows:

$$\begin{pmatrix} \delta_{11} & \dots & \delta_{j1} & \dots & \delta_{J1} \\ \delta_{12} & \dots & \delta_{j2} & \dots & \delta_{J2} \\ \dots & \dots & \dots & \dots & \dots \\ \delta_{1K} & \dots & \delta_{jK} & \dots & \delta_{JK} \end{pmatrix}$$

- $n1$ and $n0$ are the size of the two groups (by default, 100 for each group).
- *gamma* is the minimal clinically important difference (MCID) or the expected value of the difference between the means of two groups on the latent trait .
- *var* is the expected values of the variances of the latent trait (1 by default): If this option contains only one value, variances are considered to be equal between the two different groups; if this option contains two values, variances are considered to be unequal between the two groups.
- *method* allows defining the method used by the algorithm (GH, mean, mean + GH, or population + GH). By default, the method is set at GH when there is a moderate number of possible response patterns (< 1000 per group) and at population + GH otherwise.
- *nbpatterns* allows for defining the number of response patterns for which the probability of occurrence will be estimated using GM method in the mean + GH and population + GH methods. The number of response patterns that will be analyzed for each group is defined as the number of individuals in the corresponding group multiplied by *nbpatterns*. By default, *nbpatterns* is set at 2.
- *expectedpower* allows for searching for a sample size to reach a fixed level of power (the obtained sample sizes take into account the ratio between $n0$ and $n1$).

An example of syntax is given in the succeeding paragraphs; it was used in the data example section of the paper:

```
. matrix diff=(-0.328,-0.811,0.329\0.556,0.818,1.409\1.394,1.049,1.288\0.560,
              0.363,0.950)
. Raschpower, d(diff) n0(167) n1(205) gamma(0.178) var(0.77) exp(0.8)

Number of individuals in the first group: 167
Number of individuals in the second group: 205
Group effect: .178
Variance of the latent trait in the first group: .77
Variance of the latent trait in the second group: .77
Number of items: 4
Difficulty parameters of the items:

         item1  item2  item3  item4
delta_1  -.328   .556  1.394    .56
delta_2  -.811   .818  1.049    .363
delta_3   .329  1.409  1.288    .95

10%..20%..30%..40%..50%..60%..70%..80%..90%..100%

-------------------------------------------------------------------
Estimation of the variance of the group effect             0.0125
Estimation of the power                                    0.3572
-------------------------------------------------------------------
Number of patients for a power of 80.00%           517.17/ 634.85
```

The Raschpower module can be installed from Stata by using the *ssc install raschpower* command.

Table I. Estimated values of variance of the γ parameter ($Var(\hat{\gamma})$) and power ($1-\beta$) obtained with population + Gauss–Hermite methods or using simulated datasets (S_1 is the case where the variance of the latent trait and the difficulty parameters are fixed; S_2 if the case where only the difficulty parameters are fixed but the variance of the latent trait is estimated; S_3 is the case where the variance of the latent trait and the difficulty parameters are estimated) for a set of items containing three response categories, as a function of the sample size per group (N), the number of items (J), and the difference of the means between the two groups (γ).

N	J	γ	$Var_{POP}(\hat{\gamma})$	$1-\beta_{POP}$	$Var_{S1}(\hat{\gamma})$	$1-\beta_{S1}$	$Var_{S2}(\hat{\gamma})$	$1-\beta_{S2}$	$Var_{S3}(\hat{\gamma})$	$1-\beta_{S3}$
50	5	0.2	0.0598	0.127	0.0596	0.132	0.0598	0.138	0.0615	0.140
		0.5	0.0603	0.531	0.0598	0.532	0.0596	0.562	0.0627	0.562
		0.8	0.0606	0.902	0.0603	0.899	0.0608	0.898	0.0653	0.898
	10	0.2	0.0543	0.135	0.0500	0.129	0.0498	0.172	—	—
		0.5	0.0552	0.567	0.0501	0.618	0.0497	0.570	—	—
		0.8	0.0573	0.950	0.0503	0.946	0.0500	0.948	—	—
100	5	0.2	0.0299	0.211	0.0298	0.217	0.0298	0.180	0.0307	0.180
		0.5	0.0300	0.823	0.0299	0.844	0.0302	0.836	0.0314	0.836
		0.8	0.0302	0.996	0.0301	0.998	0.0302	0.998	0.0320	0.998
	10	0.2	0.0265	0.232	0.0250	0.254	0.0249	0.264	—	—
		0.5	0.0269	0.862	0.0250	0.865	0.0251	0.918	—	—
		0.8	0.0273	0.998	0.0252	0.999	0.0251	1.000	—	—
200	5	0.2	0.0149	0.374	0.0149	0.388	0.0149	0.406	0.0150	0.406
		0.5	0.0150	0.983	0.0149	0.985	0.0151	0.980	0.0154	0.980
		0.8	0.0151	1.000	0.0151	1.000	0.0152	1.000	0.0159	1.000
	10	0.2	0.0130	0.418	0.0125	0.420	0.0126	0.464	—	—
		0.5	0.0131	0.992	0.0125	0.993	0.0124	0.988	—	—
		0.8	0.0133	1.000	0.0126	1.000	0.0126	1.000	—	—
300	5	0.2	0.0099	0.519	0.0099	0.514	0.0099	0.504	0.0100	0.506
		0.5	0.0100	0.999	0.0100	0.999	0.0100	1.000	0.0102	1.000
		0.8	0.0100	1.000	0.0100	1.000	0.0102	1.000	0.0107	1.000
	10	0.2	0.0086	0.578	0.0083	0.609	0.0083	0.586	—	—
		0.5	0.0087	1.000	0.0083	1.000	0.0084	1.000	—	—
		0.8	0.0087	1.000	0.0084	1.000	0.0084	1.000	—	—
500	5	0.2	0.0060	0.736	0.0060	0.744	0.0060	0.758	0.0060	0.758
		0.5	0.0060	1.000	0.0060	1.000	0.0060	1.000	0.0061	1.000
		0.8	0.0060	1.000	0.0060	1.000	0.0061	1.000	0.0063	1.000
	10	0.2	0.0051	0.799	0.0050	0.782	0.0050	0.788	—	—
		0.5	0.0051	1.000	0.0050	1.000	0.0050	1.000	—	—
		0.8	0.0052	1.000	0.0050	1.000	0.0051	1.000	—	—

3. Results

Tables I, II, and III present the variances and the power obtained using the GH and population + GH method and for the simulated datasets (three designs S_1, S_2, and S_3) as a function of the number of individuals per group (N), number of items (J) and differences between the two means (γ) for items having $K = 3$, 5, or 7 response categories, respectively. The results of the mean and mean + GH methods are not presented here because they produced more differences with the simulation study (considered as the reference) than the population + GH method and displayed a larger computation time. GH and population + GH methods produce similar results with $J = 5$ whatever the values of the other parameters. However, the GH method could be very long to run as soon as the number of items, and the number of response categories per item were large. As a consequence, the results for the GH method are not presented in the tables for $J = 10$. Finally, for computing time issues, simulations with $J = 10$ and S3 simulation design (variance of the latent trait and difficulty parameters jointly estimated) have not been performed.

The values of $Var_{POP}(\hat{\gamma})$ and $Var_S(\hat{\gamma})$ (whatever the design S_1, S_2, or S_3) follow expected trends. The values decrease with N, J, and K and are stable whatever the value of γ. Consequently, the power follows similar trends. It increases with N, J, K, and only slightly with γ.

Table II. Estimated values of variance of the γ parameter ($Var(\hat{\gamma})$) and power ($1-\beta$) obtained with population + Gauss–Hermite methods or using simulated datasets (S_1 is the case where the variance of the latent trait and the difficulty parameters are fixed; S_2 if the case where only the difficulty parameters are fixed but the variance of the latent trait is estimated; S_3 is the case where the variance of the latent trait and the difficulty parameters are estimated) for a set of items containing five response categories, as a function of the sample size per group (N), the number of items (J), and the difference of the means between the two groups (γ).

N	J	γ	$Var_{\text{POP}}(\hat{\gamma})$	$1-\beta_{\text{POP}}$	$Var_{S1}(\hat{\gamma})$	$1-\beta_{S1}$	$Var_{S2}(\hat{\gamma})$	$1-\beta_{S2}$	$Var_{S3}(\hat{\gamma})$	$1-\beta_{S3}$
50	5	0.2	0.0509	0.142	0.0492	0.153	0.0495	0.148	0.0521	0.146
		0.5	0.0512	0.598	0.0494	0.631	0.0498	0.632	0.0532	0.634
		0.8	0.0524	0.937	0.0497	0.946	0.0504	0.950	0.0562	0.946
	10	0.2	0.0508	0.142	0.0448	0.150	0.0448	0.138	—	—
		0.5	0.0523	0.590	0.0449	0.679	0.0441	0.636	—	—
		0.8	0.0550	0.962	0.0451	0.962	0.0449	0.954	—	—
100	5	0.2	0.0250	0.244	0.0246	0.231	0.0245	0.232	0.0252	0.232
		0.5	0.0252	0.883	0.0247	0.887	0.0247	0.900	0.0260	0.900
		0.8	0.0254	0.999	0.0248	0.999	0.0249	0.998	0.0272	0.998
	10	0.2	0.0248	0.245	0.0224	0.314	0.0224	0.252	—	—
		0.5	0.0255	0.880	0.0224	0.892	0.0224	0.926	—	—
		0.8	0.0262	0.999	0.0225	1.000	0.0228	0.998	—	—
200	5	0.2	0.0124	0.435	0.0123	0.449	0.0122	0.428	0.0124	0.430
		0.5	0.0124	0.994	0.0123	0.995	0.0123	0.995	0.0128	0.995
		0.8	0.0125	1.000	0.0124	1.000	0.0125	1.000	0.0135	1.000
	10	0.2	0.0122	0.441	0.0112	0.470	0.0113	0.498	—	—
		0.5	0.0124	0.994	0.0112	0.999	0.0112	1.000	—	—
		0.8	0.0126	1.000	0.0113	1.000	0.0113	1.000	—	—
300	5	0.2	0.0082	0.597	0.0075	0.592	0.0082	0.620	0.0083	0.618
		0.5	0.0083	1.000	0.0082	1.000	0.0082	1.000	0.0085	1.000
		0.8	0.0083	1.000	0.0083	1.000	0.0084	1.000	0.0090	1.000
	10	0.2	0.0081	0.606	0.0075	0.635	0.0075	0.632	—	—
		0.5	0.0082	1.000	0.0075	1.000	0.0075	1.000	—	—
		0.8	0.0083	1.000	0.0075	1.000	0.0076	1.000	—	—
500	5	0.2	0.0049	0.813	0.0049	0.803	0.0049	0.832	0.0050	0.832
		0.5	0.0049	1.000	0.0049	1.000	0.0049	1.000	0.0051	1.000
		0.8	0.0050	1.000	0.0050	1.000	0.0050	1.000	0.0054	1.000
	10	0.2	0.0048	0.824	0.0045	0.850	0.0045	0.846	—	—
		0.5	0.0048	1.000	0.0045	1.000	0.0045	1.000	—	—
		0.8	0.0049	1.000	0.0045	1.000	0.0045	1.000	—	—

The population + GH method sometimes produces overestimated values of $Var(\hat{\gamma})$ compared with simulations, in particular for small sample sizes ($N = 50$ or 100), a large number of items ($J = 10$), and large numbers of response categories ($K = 5$ or 7). For example, the overestimation of $Var_{POP}(\hat{\gamma})$ compared with the variance obtained by simulations $Var_{S_s}(\hat{\gamma})$, $s = 1, 2, 3$ is approximately 0.0080 for $N = 50$, $J = 10$, and $K = 7$ (whatever the value of γ) versus an overestimation of only 0.0005 for $N = 500$, $J = 10$, and $K = 7$.

As a consequence of the overestimation of $Var_{S_s}(\hat{\gamma})$, $s = 1, 2, 3$, the power estimated with the population + GH method can be underestimated compared with the power obtained with the simulations (S1 design). This underestimation can reach 0.089 in the worst case (for $N = 50$, $J = 10$, $\gamma = 0.5$, and $K = 5$); nevertheless, it is only 0.010 on average. We note that the underestimation of the power is mainly related to scenarios with a small sample size ($N = 50$) and a large number of possible responses patterns (when $J = 10$), especially when the power has a medium value (between 0.3 and 0.7).

Compared with the S1 design, the S2 and S3 designs (variance of the latent trait and difficulty parameters jointly estimated with the group effect) tend to produce somewhat larger estimations than $Var_{S1}(\hat{\gamma})$, but this effect seems to be relatively negligible on the corresponding powers, which remain close to one another.

Table III. Estimated values of variance of the γ parameter ($Var(\hat{\gamma})$) and power ($1-\beta$) obtained with population + GH methods or using simulated datasets (S_1 is the case where the variance of the latent trait and the difficulty parameters are fixed; S_2 if the case where only the difficulty parameters are fixed but the variance of the latent trait is estimated; S_3 is the case where the variance of the latent trait and the difficulty parameters are estimated) for a set of items containing seven response categories, as a function of the sample size per group (N), the number of items (J), and the difference of the means between the two groups (γ).

N	J	γ	$Var_{POP}(\hat{\gamma})$	$1-\beta_{POP}$	$Var_{S1}(\hat{\gamma})$	$1-\beta_{S1}$	$Var_{S2}(\hat{\gamma})$	$1-\beta_{S2}$	$Var_{S3}(\hat{\gamma})$	$1-\beta_{S3}$
50	5	0.2	0.0498	0.144	0.0462	0.137	0.0466	0.154	0.0495	0.152
		0.5	0.0503	0.606	0.0464	0.636	0.0463	0.642	0.0500	0.640
		0.8	0.0508	0.944	0.0467	0.963	0.0471	0.966	0.0536	0.0962
	10	0.2	0.0511	0.141	0.0433	0.175	0.0429	0.162	—	—
		0.5	0.0530	0.584	0.0433	0.654	0.0434	0.670	—	—
		0.8	0.0549	0.927	0.0435	0.957	0.0435	0.972	—	—
100	5	0.2	0.0243	0.250	0.0231	0.252	0.0232	0.250	0.0240	0.250
		0.5	0.0243	0.894	0.0232	0.898	0.0233	0.918	0.0247	0.916
		0.8	0.0247	0.999	0.0234	1.000	0.0234	0.994	0.0256	0.994
	10	0.2	0.0249	0.244	0.0216	0.271	0.0215	0.242	—	—
		0.5	0.0256	0.878	0.0217	0.897	0.0216	0.920	—	—
		0.8	0.0263	0.999	0.0218	0.999	0.0218	1.000	—	—
200	5	0.2	0.0119	0.451	0.0116	0.471	0.0116	0.478	0.0119	0.476
		0.5	0.0119	0.996	0.0116	0.998	0.0116	0.996	0.0121	0.996
		0.8	0.0120	1.000	0.0117	1.000	0.0117	1.000	0.0128	1.000
	10	0.2	0.0122	0.442	0.0108	0.495	0.0108	0.466	—	—
		0.5	0.0125	0.994	0.0108	0.999	0.0109	0.994	—	—
		0.8	0.0127	1.000	0.0109	1.000	0.0108	1.000	—	—
300	5	0.2	0.0078	0.618	0.0077	0.619	0.0077	0.666	0.0078	0.668
		0.5	0.0079	1.000	0.0077	1.000	0.0078	1.000	0.0081	1.000
		0.8	0.0079	1.000	0.0078	1.000	0.0078	1.000	0.0084	1.000
	10	0.2	0.0080	0.610	0.0072	0.646	0.0072	0.644	—	—
		0.5	0.0082	1.000	0.0072	1.000	0.0072	1.000	—	—
		0.8	0.0083	1.000	0.0073	1.000	0.0073	1.000	—	—
500	5	0.2	0.0047	0.833	0.0046	0.837	0.0046	0.836	0.047	0.836
		0.5	0.0047	1.000	0.0046	1.000	0.0047	1.000	0.048	1.000
		0.8	0.0047	1.000	0.0047	1.000	0.0047	1.000	0.051	1.000
	10	0.2	0.0047	0.830	0.0043	0.855	0.0043	0.860	—	—
		0.5	0.0048	1.000	0.0043	1.000	0.0043	1.000	—	—
		0.8	0.0049	1.000	0.0044	1.000	0.0044	1.000	—	—

4. Example of determination of the power to compare two groups of pathological gamblers

To illustrate this approach in the epidemiological research, we use data related to a cohort of 628 gamblers. In this cohort, gamblers are recruited in gambling places or by press. Problematic gamblers are defined as those obtaining more than three criteria from the gambling section of the DSM – Fourth Edition. In this cohort, 372 gamblers are considered problematic. Among these problematic gamblers, 167 seek a treatment for this addiction, and 205 are not treatment seeking. At the baseline, the gamblers of this cohort respond to several questionnaires and in particular to the Gambling Attitudes and Beliefs Survey [24]. This questionnaire is composed of 35 items having four response categories ('strongly agree', 'agree', 'disagree', 'strongly disagree') and explores gambling-related dysfunctional beliefs. In this questionnaire, four items allow computing a score based on dysfunctional beliefs about luck [25] (items 8 'Some people are unlucky', 12 'Some people are lucky to have around when I'm gambling', 20 'I have carried a lucky charm when I gambled', and 29 'Some people can bring bad luck to other people'). An assumption is that treatment for gambling addiction allows reducing these dysfunctional beliefs, in particular the one related to the impact of luck on gambling. The four items corresponding to these dysfunctional belief dimensions follow, in this sample, a PCM ($p = 0.96$, goodness of fit Glas' R1m test [26]).

A PCM includes a group variable to distinguish problematic gamblers seeking a treatment from the others was fitted. The group effect was estimated at 0.178, which is the difference between the means of these two groups ($p = 0.119$). Because the variance of the latent trait is estimated at 0.77, the standardized difference in the latent trait scale ($0.178/\sqrt{0.77} \approx 0.20$) between the two groups can be considered small [27]. Nevertheless, this difference corresponded to an average difference of 2.0 on the score scale, and this difference has been considered clinically relevant by psychiatrists. Because the test of the comparison of the two groups is not significant, it is interesting to evaluate its power. The matrix of estimated difficulty parameters is the following:

$$\begin{pmatrix} -0.328 & 0.556 & 1.394 & 0.560 \\ -0.811 & 0.818 & 1.049 & 0.363 \\ 0.329 & 1.409 & 1.288 & 0.950 \end{pmatrix}$$

Under these conditions, the Raschpower Stata module evaluates the power of the test at 35.7%. To attain a power of 80%, the required sample size would be 582 patients per group (or 518 and 635 patients in the first and second group, respectively, to take into account the ratio between problematic gamblers with treatment and problematic gamblers without treatment).

5. Discussion

The Raschpower procedure was initially developed for dichotomous items. It allows for estimating the power of a Wald test to compare the difference of the means of two groups of patients on a latent variable measured by a Rasch model. This method consists in defining a planning dataset built from the probability of observing each response pattern using GM method. From this planning dataset, the difference between the means of the two groups is estimated with its variance, and the power of the test can then be evaluated.

In this paper, this method has been adapted to polytomous items, which are more often encountered in clinical research. For a small number of possible response patterns (up to several thousand), no adaptation is necessary. It is possible to estimate the probabilities of each response pattern using GM method; consequently, a planning dataset can be obtained. However, when the number of items and response categories for each item increases, the number of possible response patterns becomes very large (up to several hundred million); it is not possible in this case to estimate the probability of observing each of them. Hence, for such a situation, we propose the population + GH method. This method is based on the simulation of one million individuals, and the probabilities associated with the most frequently observed responses patterns are estimated.

In the present paper, results obtained by the GH and population + GH methods are compared with results obtained by a simulation study in some practical situations: five or 10 items, with three, five, or seven response categories; samples of 50 to 500 individuals per group; and for various values of the difference between the means of the two groups (0.2, 0.5, or 0.8) with three designs of analysis for the simulated dataset corresponding to the fact that the variance of the latent trait was either fixed (S1) or not (S2 and S3) and that the difficulty parameters were either fixed (S1 and S2) or not (S3). In all of these cases, the population + GH method allows for estimating the power with an acceptable precision compared with the simulations (on average, a difference of 0.010 is observed with a weak quasi-systematic underestimation of the power for the population + GH method compared with the following simulations: 85 cases among the 90 tested with S1, 81 cases among 90 tested with S2, with a maximum difference estimated at 0.089).

Compared with the simulations, GH has the advantage of being a deterministic method and is also faster when the number of possible responses patterns is moderate. Population + GH is a useful alternative to GH when the number of possible response patterns becomes very large (it avoids estimating the probability of each of them). Even if population + GH is not strictly a deterministic method, its results seem to be very stable if the population size is very large compared with the expected samples; in the present paper, a population of one million individuals has been used for samples of 50 to 500 individuals per group.

The main advantage of GH and its alternative population + GH method is the ability to obtain a reliable approximation of the power of a Wald test to compare the means of two groups on a latent variable measured by a Rasch family model. Empirically, it can be demonstrated that the population + GH method runs faster than GH when the number of possible response patterns per group is approximately greater than 1000. Consequently, we can recommend using the population + GH method when the scale is

composed of more than 10 dichotomous items or six items with three response categories, four items with five response categories, or when the number of response categories is greater than five.

Such an estimation of the power is very useful during the planning step of clinical or epidemiological studies, to define a required sample size in a large number of situations often encountered in practice or during the analysis of a dataset to determine the post hoc power of a test (especially if the result is not significant). In this latter case, all of the required parameters are available (difficulty parameters, means of the latent variable in the two groups, and variance of the latent variable), and it is easy to obtain the estimation of the post hoc power. When planning a study, expected values of these parameters must be defined, using pilot studies or assumptions; misspecifications of these values can create a poor estimation of the power and consequently either an overestimation or an underestimation of the corresponding sample size. The impact of the misspecification of the parameters in the planning step should be investigated. For example, difficulty parameters can be estimated from pilot studies, but such estimations can have poor reliability and consequently may lead to improper values. However, a small impact of a weak or medium misspecification of difficulty parameters has been demonstrated on power for the comparison of two groups of patients with a Rasch model on power [17]. We might expect similar results with the PCM, but it has to be investigated. Moreover, the impact of critical situations (in terms of the distribution of the item difficulty) on the results of the Raschpower procedure has been investigated [28] and has exhibited a good robustness for the Raschpower procedure to the specific distribution of the difficulty parameters.

An additional misspecification could be considered for the assumed value of the variance of the latent trait. It can be difficult to obtain a reliable estimation of this parameter from pilot studies. We may suspect an important impact of this misspecification: For example, in our cohort of gamblers, the standard error of the estimation of the variance parameter is 0.13, and consequently, the 95% confidence interval of the variance parameter is [0.47; 1.03]. Using these two bounds of the confidence interval, the power of the test varied between 29.6% and 46.9% (estimated with Raschpower), and in terms of the required sample size for an expected power of 80%, we obtain a number between 414 and 722 gamblers per group. The uncertainty in the variance estimation may lead to a substantial difference in sample size estimation in the context of clinical or epidemiological research.

We note that the power computed using the Raschpower algorithm is based on the idea that the variance of the latent trait (and the difficulty parameters) are considered known parameters and are not estimated using a dataset. This idea is similar to the classical approaches to determine power and sample size using manifest variables [14]. S2 and S3 designs for the simulation study allow being closer to a practical situation, by relaxing these assumptions on the variance of the latent trait and on the difficulty parameters. These designs generally produce a small increase in the variance of the γ estimator, but this increase has a negligible impact on the estimation of power. This result is close to [29] for manifest variables, with a negligible impact of the estimation of the variance on power as soon as the variance was estimated with an important number of degrees of freedom (variation of the required sample size was lesser than 4% for a variance estimated with 100 degrees of freedom for example).

Finally, the average difference between the two groups (γ) can be considered in two ways, according to the step at which the power is computed. In the planning step, it is recommended to use the MCID if it is available. As a consequence, there is no 'misspecification' of this parameter because its value is independent of the sample and only relies on the measure (that is to say, provided by the questionnaire). Nevertheless, it could be difficult to define it on the latent trait [30] because a latent trait is more difficult to conceptualize for a physician (or other users of such methods) than a manifest variable. When analyzing data, the observed difference between the two means is often used instead of the MCID to compute power (this is the case in our illustration). In this case, we can suspect misspecifications for this parameter. For example, in our dataset, the 95% confidence interval of the difference between means is $[-0.046; 0.402]$. With these two bounds, the power of the test varied from 6.1% to 94.8%.

However, the impact of making wrong assumptions regarding the expected model on sample size and power determination is also of interest. For example, the violation of some assumptions of the IRT model could be investigated, such as non-normality of the latent variable, non-respect of the local independence or of the unidimensionality. Indeed, in theory, these properties should be investigated during the validation of the scales, but in practice, the violation of some of these assumptions is possible. In dichotomous situations, non-respect of the normality assumption of the latent trait and non-respect of the local independence have been investigated [31, 32]; this allows the conclusion of a negligible impact of the violation of these assumptions, but these investigations should be extended to the polytomous case.

Extending the proposed approach to other designs often used with PRO data, such as longitudinal studies, would also be worthwhile. Longitudinal IRT models 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 (as realized by [33]), or a group effect in the case of a two-sample design). The latter case should be studied in future research.

As a conclusion, the proposed approach for polytomous Rasch models is based on a numerical approach to estimate the standard error of the group effect and produces satisfying results to evaluate the power of a group comparison test in this framework. Nevertheless, this procedure could be improved in terms of computation time by using analytical approaches based on the property of sufficiency of the score for the latent trait of the Rasch models, as proposed by [34].

Acknowledgments

This study was supported by the French National Research Agency, under reference N 2010 PRSP 008 01.

References

1. Mohammed H, Kaufman C, Limbrick D, Steger-May K, Grubb R, Rothman SL, Weisenberg J, Munro R, Smyth M. Impact of epilepsy surgery on seizure control and quality of life: a 26-year follow-up study. *Epilepsia* 2012; **53**(4):712–720.
2. Strand V, Smolen J, van Vollenhoven R, Mease P, Burmester G, Hiepe F, Khanna D, Nika E, Coteur G, Schiff M. Certolizumab pegol plus methotrexate provides broad relief from the burden of rheumatoid arthritis: analysis of patient-reported outcomes from the RAPID 2 trial. *Annals of the rheumatic diseases* 2011; **70**(6):996–1002.
3. Gotay C, Kawamoto C, Bottomley A, Efficace F. The prognostic significance of patient-reported outcomes in cancer clinical trials. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 2008; **26**(8):1355–1363.
4. Smith E, Pang H, Cirrincione C, Fleishman S, Paskett E, Ahles T, LR Bressler, Fadul C, Knox C, Le-Lindqwister N, Gilman PB, Shapiro CL. Alliance for clinical trials in oncology: effect of duloxetine on pain, function, and quality of life among patients with chemotherapy-induced painful peripheral neuropathy: a randomized clinical trial. *JAMA* 2013; **309**: 1359–1367.
5. Lamy A, Devereaux P, Prabhakaran D, Taggart D, Hu S, Paolasso E, Straka Z, Piegas LS, Akar A, Jain A, Noiseux N, Padmanabhan C, Bahamondes JC, Novick RJ, Vaijyanath P, Reddy SK, Tao L, Olavegogeascoechea PA, Airan B, Sulling RP, Ou Y, Pogue J, Chrolavicius S, Yusuf S. Coronary investigators: effects of off-pump and on-pump coronary-artery bypass grafting at 1 year. *New England Journal of Medicine* 2013; **368**:1179–1188.
6. Cunningham MA, Swanson V, Holdsworth RJ, O'Carroll RE. Late effects of a brief psychological intervention in patients with intermittent claudication in a randomized clinical trial. *The British Journal of Surgery* 2013; **100**:756–760.
7. Cartwright M, Hirani S, Rixon L, Beynon M, Doll H, Bower P, Bardsley M, Steventon A, Knapp M, Henderson C, Rogers A, Sanders C, Fitzpatrick R, Barlow J, Newman SP. Whole systems demonstrator evaluation team: effect of telehealth on quality of life and psychological outcomes over 12 months (whole systems demonstrator telehealth questionnaire study): nested study of patient reported outcomes in a pragmatic, cluster randomised controlled trial. *BMJ* 2013; **346**:f653.
8. Jose A, Olino T, O'Leary K. Item response theory analysis of intimate-partner violence in a community sample. *Journal of Family Psychology: JFP: Journal of the Division of Family Psychology of the American Psychological Association (Division 43)* 2012; **26**(2):198–205.
9. Olino TM, Yu L, Klein DN, Rohde P, Seeley JR, Pilkonis PA, Lewinsohn PM. Measuring depression using item response theory: an examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research* 2012; **21**(1):76–85.
10. Rasch G. *Probabilistic models for some intelligence and attainment tests* expanded edition. The University of Chicago Press: Chicago, 1980.
11. Masters G. A Rasch model for partial credit scoring. *Psychometrika* 1982; **47**(2):149–174. http://ideas.repec.org/a/spr/psycho/v47y1982i2p149-174.html.
12. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, Brédart 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: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 2004; **13**(10):1683–1697.
13. Cella D, Beaumont J, Webster K, Lai J, Elting L. Measuring the concerns of cancer patients with low platelet counts: the functional assessment of cancer therapy – thrombocytopenia (FACT-Th) questionnaire. *Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer* 2006; **14**(12):1220–1231.
14. Julious S. *Sample Sizes for Clinical Trials*. CRC Press - Taylor & Francis: Boca Raton, Florida, 2009.
15. Schmoor C, Sauerbrei W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine* 2000; **19**(4):441–452.
16. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology* 1991; **44**(8):763–770.
17. 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.
18. Hardouin JB, Amri S, Feddag M, Sébille V. Towards power and sample size calculations for the comparison of two groups of patients with item response theory models. *Statistics in Medicine* 2012; **31**(11-12):1277–1290.

19. Mislevys R. Estimating latent distributions. *Psychometrika* 1984; **49**:359–381.
20. Mislevys R. Estimation of latent group effects. *Journal of the American Statistical Association* 1985; **80**:993–997.
21. Mislevys R. Randomization-based inference about latent variables from complex samples. *Psychometrika* 1991; **56**: 177–196.
22. Hamel J, Hardouin JB, Le Nel T, Kubis G, Roquelaure Y, Sébille V. Study of different methods for comparing groups by analysis of subjective health measurements. *PloS One* 2012; **7**(10):e44695.
23. Christensen K, Bjorner J, Kreiner S, Petersen J. Latent regression in loglinear rasch models. *Communications in Statistics: Theory and Methods* 2004; **33**:1295–1313.
24. Breen R, Zuckerman M. 'Chasing' in gambling behavior: personality and cognitive determinants. *Personality and Individual Differences* 1999; **27**:1097–1111.
25. Bouju G, Hardouin JB, Boutin C, Gorwood P, Le Bourvellec J, Feuillet F, Venisse J, Grall-Bronnec M. A shorter and multidimensional version of the gambling attitudes and beliefs survey (GABS-23). *Journal of Gambling Studies* 2014; **30**(2):349–367.
26. Glas C. The derivation of some tests for the rasch model from the multinomial distribution. *Psychometrika* 1988; **53**(4): 525–546.
27. Cohen J. *Statistical Power Analysis for the Behavioral Sciences (Second Ed.)* Lawrence Erlbaum Associates: Mahwah, New Jersey, 1988.
28. Blanchin M, Hardouin JB, Guillemin F, Falissard B, Sébille V. Power and sample size determination for the group comparison of patient-reported outcomes with Rasch family models. *PloS One* 2013; **8**(2):e57279. DOI: 10.1371/journal.pone.0057279.
29. Julious S, Owen R. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics* 2006; **5**:29–37.
30. Rouquette A, Blanchin M, Sébille V, Guillemin F, Côté S, Falissard B, Hardouin JB. Determination of the minimal clinically important difference using item response theory models: an attempt to solve the issue of the association with baseline score. *Journal of Clinical Epidemiology* 2014; **4**(64):433–440.
31. Guilleux A, Blanchin M, Hardouin JB, Sébille V. Power and sample size determination in the rasch model: evaluation of the robustness of a numerical method to non-normality of the latent trait. *PloS one* 2013; **9**(1):e83652.
32. Feddag M, Sébille V, Blanchin M, Hardouin JB. Estimation of parameters of the Rasch model and comparison of groups in presence of locally dependent items. *Journal of Applied Measurement* 2015. under press.
33. Feddag M, Blanchin M, Hardouin JB, Sébille V. Power analysis on the time effect for the longitudinal Rasch model. *Journal of Applied Measurement* 2014; **3**(15):292–301.
34. Glas C, Verhelst N. Extensions of the partial credit model. *Psychometrika* 1989; **54**:635–659.