Tutorials in Quantitative Methods for Psychology
2010, Vol. 6(2), p. 39-51.

DOI: 10.20982/tqmp.06.2.p039

Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation

Marlène Galdin
Centre de réadaptation en déficience intellectuelle et troubles envahissants du développement Mauricie/Centre-du-Québec - Institut universitaire

Louis Laurencelle
Université du Québec à Trois-Rivières

Statistical properties of the ability level estimate (θ) in item response theory (IRT) were investigated through a Monte Carlo investigation, based on data generated with a four cases multifactor design. Dichotomous items and the logistic two-parameter IRT model in a one-dimensional setting have been chosen. The estimation procedure was the marginalized Bayesian item parameters estimation and EAP estimation for θ. The property of invariance is discussed. Results show that estimation of θ is intrinsically biased, is constrained by the number of items and that it performs better when the number of items and the number of examinees increase. Furthermore, IRT parameters do not seem to perform better nor give more information than those used in classical test theory.

Classical test theory (Gulliksen, 1950; Lord & Novick, 1968; Laveault & Grégoire, 2002) proposes an algebraic-conceptual framework to explore the connection between an observed score measured by a test which evaluates a skill, knowledge or psychological aptitude, and the person's unknown true score or ability level. Item response theory (IRT) (Hambleton, Swaminathan & Rogers, 1991; Bertrand & Blais, 2004), on the other hand, tackles the same problem on a molecular basis, i.e. item-wise, by trying to model the interaction between the respondent's ability level and the operational characteristics of each item. An attractive feature of IRT is its parametric setting, usually represented with a *k*-item parameter logistic probability model (*k* = 1, 2, 3), and the property of invariance associated with it (McKinley, 1989).

The full 3-item parameter logistic model serves to illustrate the role and interpretation of each component: it describes the examinee's probability of giving the correct response to an item:

$$P_j(\theta_r) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta_r - b_j)}} \qquad (1)$$

In equation (1), θ_r denotes r^{th} examinee's ability level, b_j is the item's difficulty level, a_j its coefficient of discrimination at the inflexion point, and c_j the index of pseudo-guessing. Values for θ and b_j range currently from –3 to 3, a_j is usually a small positive value, and c_j, varying from 0 to 1, is used mostly for multiple-choice items where chance supplies a minimum probability of guessing the correct answer or the probability of low level respondents to obtain the correct answer. The 2-parameter model does away with the c_j parameter (i.e. $c_j \equiv 0$) and the 1-parameter model (e.g. Rasch

M. G.: Centre de réadaptation en déficience intellectuelle et troubles envahissants du développement Mauricie/Centre-du-Québec - Institut universitaire, 3090, rue Foucher Trois-Rivières (Qc) G8Z1M3; Email : marlene_galdin@ssss.gouv.qc.ca; L. L.: Département des sciences de l'activité physique, Université du Québec à Trois-Rivières.
We wish to thank the reviewers for their carefull revision and the suggestions they made. They greatly enhanced the quality of this paper.

model) uses only the b_j coefficient (e.g. setting $a_j \equiv 1$).

Item response theory enhances and in some way supplants classical test theory (CTT) by implementing new concepts and a new vocabulary to describe tests (item characteristic curve, item/test information function, optimal testing, etc.) and by putting the focus on the estimation of items' operational characteristics (e.g. assessment of test dimensionality, estimation of the *a*, *b*, *c* parameters, item bias and differential item functioning), although these issues are also addressed in CTT. Moreover, tenants of IRT put forward the property of invariance possessed by parameter estimates, advocating that such estimates, that of θ for instance, are obtained free of context and can be deemed truly characteristic of their object, by opposition to the context-bound estimates in CTT. "Invariance" often means that values of IRT item parameters ought to be identical for separate groups of examinees and through different measurement conditions (Rupp & Zumbo, 2006).

What is invariance? Like most authors on the same topic, Hambleton et al. (1991) stress the importance of this concept as a distinctive asset of IRT:

> The importance of the property of invariance of item and ability parameters cannot be overstated. This property is the cornerstone of item response theory and makes possible such important applications as equating, item banking, investigation of item bias, and adaptive testing. (p. 25)

On the one hand, "invariance" means equality: "If invariance holds, the parameters obtained should be identical" (Hambleton et al., 1991, p. 20; Rupp & Zumbo, 2006, p. 64). On the other hand, a less stringent form of correspondence, e.g. linear equivalence, is admitted as a demonstration of invariance: two sets of parameters are said mutually "invariant" if they may be linearly transformed one into the other (Hambleton et al., 1991; Rupp & Zumbo, 2006; Stocking and Lord, 1983)[1]. This second meaning of "invariance", also named "congruence", is akin to the notion of (linear) correlation, to the point that values of Pearson's correlation coefficients are taken as conclusive indications of invariance (Fan, 1998; Frenette, et al., 2007), with a threshold value of $r = 0.90$ being proposed.

From another standpoint, that of estimation theory in mathematical statistics (Kendall & Stuart, 1977; Freund, 1992), the concept of invariance must be translated into affine concepts, notably the concept of "bias". An estimating function based on a random sample of a population is said to be unbiased if its expectancy (across samples) is equal to the target parametric value. For the ability parameter of respondent "*r*", this simply means:

$$\mathbf{E}\{\hat{\theta}_r\} = \theta_r \qquad (2)$$

concurring with the "identical" definition of invariance in Hambleton et al. (1991), the bias being measured by the difference between $E\{\hat{\theta}_r\}$ and θ_r, here $E\{\hat{\theta}_r\} - \theta_r = 0$.

As Mckinley (1989) pointed out, the first step before using an IRT model is to estimate its parameters; usually none of them are known a priori. Baker and Kim (2004) provide a broad coverage of the methods and procedures for estimating the parameters of test items and examinees' ability levels. Pragmatically, the LOGIST™ program (Barton & Lord, 1982) was popular for a while; the method implemented in that program was called joint maximum likelihood estimation (JMLE) and was formulated by Birnbaum (1968): the θ and item parameters were simultaneously estimated.

Other popular programs such as BILOG-MG 3TM (Zimowski, Muraki, Mislevy & Bock, 2003) or MULTILOG V7TM (Thissen, Chen & Bock, 2003) use (optionally) the marginal maximum likelihood estimation (MMLE), and an expectation-maximization (EM) algorithm. This technique estimates the items' and θ parameters in consecutive steps. The advantage is that convergence can be reached with a fixed number of items without calling upon an arbitrary prior ability distribution. Baker and Kim (2004) recommend using the marginalized Bayesian item parameter estimation (BME). This estimation is quite similar to the MMLE, except that a prior distribution is added on the discrimination parameter (*a*). BME ensures that the procedure can be completed even in limit cases (e.g. when all items have been answered correctly or all incorrectly). Once the item parameters are "calibrated", i.e. estimated, the θ parameters are obtained by the largely used Bayes Expected A Posteriori (EAP) estimation procedure proposed by Bock and Mislevy (1982).

The lack of invariance or the so called item parameter drift has been studied by others (e.g. Frenette, et al., 2007; Rupp & Zumbo, 2006; Si & Schumacher, 2004; Wainer & Thissen, 1987; Wells, Subkoviak & Serlin, 2002). Results show that there might be a slight lack of invariance or item parameter drift under particular conditions (e.g. test length, number of examinees, presence of other latent traits), but findings are not unequivocal for specific conditions which worsen those variations. As pointed out earlier, accuracy of measurements is important in order to help users choose the best model which fits their reality or need.

The main purpose of this study was to shed more light on the invariance of estimation of θ, *b* and *a* parameters, in the context of a largely used two-phase estimation procedure. This paper presents a Monte Carlo investigation based on a four cases design which reproduces conditions that might be found in different testing contexts. Indeed, test reliability, test length, number of examinees and values of

item parameters vary widely from one context to another, so that we set up three main cases divided into four sub-cases to cover a large array of possibilities reflecting realistic conditions. To complete our investigation, a fourth case was designed based on a pioneering idea. This idea relates to the question of what might happen with regard to one's ability estimation if some other group of individuals with very different ability levels are introduced in the estimation process : does one's ability estimate keep invariant whatever group of individuals it is embedded in, or otherwise what are the effects of such a relocation? The salient questions that we wished answered were the following: How do the estimated abilities ($\hat{\theta}$) match their corresponding generated values (θ)? What factors do influence the consistency, reliability and other indicators of $\hat{\theta}$'s invariance? Do the estimated abilities ($\hat{\theta}_r$), and the more classical X scores (the sum of answers), behave equivalently across conditions, and what distinguishes $\hat{\theta}_r$? In order to provide complete information to the reader, other questions and other answers will also be examined with regard to parameters *a* and *b* and their corresponding CTT indices. The parametric organization and details of the experiment together with the modalities of the Monte Carlo implementation are laid out in the next section.

1. STUDY DESIGN

This study is a Monte Carlo investigation. Dichotomous (0/1) items and the logistic two-parameter IRT model (*a* and *b*) in a one-dimensional setting have been chosen. First, item parameters *a* and *b* are generated, followed by θ values, then, from these, random item response patterns for each examinee are generated twice, once at "pre-test" and again at "post-test". Numbers of items and examinees are given in the *Cases* section below, together with their parametric conditions. The whole procedure is iterated 30 times within each condition. Means, standard deviations and Pearson correlation coefficients are computed for varying outcome indices across the 30 iterations: the relatively low noise of our simulated data coupled with the high efficacy (cf. *F* tests and ω^2) of our independent variables obviated the need of unduly slowing down our experimentation with more replications. Data generation and compilation as well as the handling of experimental conditions were programmed in Borland's Delphi 5™ language and run on a PC computer platform.

Experimental conditions

Cases

Four basic cases have been designed. Cases 1, 2 and 3 present the same crossed experimental conditions : number of examinees (N = 100, 500, 1000), number of items (*k* = 10, 30, 50, 100) and source parameters *b* and θ generated from populations each having μ = 0, 1 and 2.

The main difference between cases 1, 2 and 3 is the reliability (ρ_{XX}) condition between the pre- and post-test *X* scores. There are reciprocal relations between the test-retest reliability (ρ_{XX}) of *X* scores, number of items (*k*) and the item discrimination coefficients (*a*) in the logistic two-parameter model, i.e. a set of *k* items having Gamma-generated *a* coefficients with mean μ_a will result in a specific mean value of ρ_{XX}^2. Cases 1, 2 and 3 are explained below. In Case 4, we introduce "witness protocols", i.e. sets of responses from a few respondents that are transferred unchanged from pre-test to post-test and are then mingled with freshly generated data, the purpose being to measure the robustness of θ estimates when the estimation environment changes.

Case 1: Reliability increasing with k (μ_a = 0.5)

A single value of hyper-parameter μ_a was used, entailing rising values of test-retest reliability for increasing number (*k*) of items, i.e. ρ_{XX} = 0.350, 0.618, 0.729 and 0.843 for *k* =10, 30, 50 and 100. All combinations of μ_b = 0, 1, 2, μ_θ = 0, 1, 2 and N = 100, 500, 1000 were used.

Case 2: Low reliability ($\rho_{XX} \approx 0.40$)

A common low value of test-retest reliability was imposed for all *k* through a lessening of μ_a, i.e. $\mu_a \approx 0.564$, 0.308, 0.236 and 0.165 for *k* = 10, 30, 50 and 100 respectively. The same combinations of μ_b, μ_θ and N were applied as in case 1.

Case 3: High reliability ($\rho_{XX} \approx 0.80$)

A common high value of test-retest reliability was imposed for all *k* through a variation of μ_a, i.e. $\mu_a \approx 1.843$, 0.858, 0.628 and 0.423 for *k* = 10, 30, 50 and 100. The same combinations of μ_b, μ_θ and N were applied as in cases 1 and 2.

Case 4: Witness response protocols

Ability (θ) estimation and the associated invariance principle of IRT being the main concerns of this study, we contrived a way of ascertaining the reliability and stability of $\hat{\theta}_r$ estimates by manipulating the sampling conditions of estimation. Thus, a "witness response protocol" is a protocol generated at pre-test which is identically reproduced at post-test, while "companion protocols" are allowed to vary, i.e. are generated afresh at post-test from their original θ parameter. In all sub-cases of case 4, only conditions with N = 500 examinees and *k* = 30 items were studied, together μ_a = 0.5, μ_b = 0 and (except case 4d) μ_θ = 0; note that the μ_a = 0.5, *k* = 30 couple entails a moderate reliability of $\rho_{XX} \approx 0.618$. As for every combination of conditions in cases 1, 2 and 3, each

sub-case of case 4 was iterated 30 times. At each iteration, ten (10) Monte Carlo runs were effected, and the outcomes and estimates for the 10 first examinees (the "witnesses") were extracted and stored, so that 100 (= 10 × 10) witness sets of data were produced per iteration and submitted to analysis. Specific conditions for each sub-case are described below.

Sub-case 4a. For this sub-case, used as a standard for comparison, the 10 witness protocols of a Monte Carlo run are in fact generated at each of pre- and post-test times. Explicitly, 500 θ-values (with $\mu_\theta = 0$) are generated once, and response protocols are generated anew at pre- and then at post-test for all examinees (control condition).

Sub-case 4b. In each run, 500 θ-values (with $\mu_\theta = 0$) are generated, along with 500 response protocols at pre-test. At post-test, the first 10 protocols of pre-test are reproduced identically as witnesses, and the remaining 490 (= 500 – 10) protocols are generated afresh ("same companions" condition).

Sub-case 4c. In each run, 500 θ-values (with $\mu_\theta = 0$) are generated, together with the 500 response protocols for pre-test. At post-test, the first 10 protocols of pre-test are reproduced identically as witnesses; for the remaining part of the sample, 490 (= 500 – 10) new θ-values (still with $\mu_\theta = 0$) are generated along with their random response protocols ("equal new companions" condition).

Sub-case 4d. In each run, 500 θ-values (with $\mu_\theta = 0$) are generated, together with the 500 response protocols for pre-test. At post-test, the first 10 protocols of pre-test are reproduced identically as witnesses ; for the remaining part of the sample, 490 (= 500 – 10) new θ-values, under hyper-parameter $\mu_\theta = 1$ and generally higher, are produced, and their corresponding random response protocols are obtained ("better new companions" condition).

Parameters generation

Generation of θ

The one-dimensional ability level (θ) was generated as a random normal deviate, with μ_θ as specified (e.g. 0, 1 or 2) and $\sigma_\theta^2 = 1$.

Generation of b

The parameter embodying item difficulty level was likewise generated as a random normal deviate, with μ_b as specified (e.g. 0, 1 or 2) and $\sigma_b^2=1$.

Generation of a

The item discrimination parameter (a), in a one-dimensional psychometric model, is typically positive and positively skewed (Baker & Kim, 2004). From practical as well as realistic considerations, we chose the χ^2-family distribution (member of the Gamma family) with index parameter (degrees of freedom) 20, whose parametric moments are $\mu = 20$, $\sigma^2 = 40$, $\gamma_1 \approx 0.632$ and $\gamma_2 \approx 0.600$. Thus, each sampled χ_{20}^2 variate was transformed to a random a coefficient through "$\chi^2 / 20 \times \mu_a$", with resulting parametric moments $\mu = 1$, $\sigma^2 = \mu_a^2 / 10$, $\gamma_1 \approx 0.632$ and $\gamma_2 \approx 0.600$. The values attributed for μ_a were discussed above.

Item parameters and ability estimation

In BILOG-MG 3™ or MULTILOG™, the default estimation procedures are the marginalized Bayesian item parameter estimation (Bayesian Modal Estimation – BME) via an EM algorithm (Dempster, Laird, & Rubin, 1977) for item parameters and the Bayes Expected A Posteriori (EAP) estimation for θ. In this study, the same procedures were used through computer freeware called Libirt (Item Response Theory Library, version 0.8.4)[3] (Germain, Valois, & Abdous, 2008). Although it was already validated, we submitted the Libirt procedure to independent checks, the EIRT estimates coinciding satisfactorily[4] with those from BILOG-MG 3™.

The default values and prior distribution needed for the a and b parameters as well as the reference distribution for θ are the same that in BILOG-MG 3™. After being generated as explained in the previous section, the response protocols were processed through the two-phase Libirt program in order to obtain item parameters and ability estimates.

For the BME/EM process, a normal prior distribution was used for the item difficulty parameter b, and a prior lognormal distribution was used for the discrimination parameter a (with $\mu_a = 1.70$ and $\sigma = 2.81$). Considering numbers of items and subjects under some conditions, for the EM algorithm, we chose to run a maximum of 100 iterations, and precociously ended when the desired precision (i.e. 10^{-5}) was achieved. In the context of the marginalization, the θ-values were assumed to follow a standard normal distribution.

In the EAP procedure, a non-iterative algorithm, each θ is individually estimated as a weighted average across the θ-domain (uniformly distributed from –4 to 4); the weighting factor is the joint probability $\prod_1^k P_j(\theta)$ for the k items using equation (1), with $c_j \equiv 0$.

Other considerations on the estimation procedures will be brought up later, in the discussion section.

2. RESULTS

In this section, we first identify the various quantities produced and recorded for this Monte Carlo investigation; statistical treatment methods are also outlined. Results pertaining to ability estimates are then examined, and finally complementary results about the estimation of item

parameters are reviewed.

Statistical data and methods

Effectuated under the experimental design, each Monte Carlo run handled a few sets of variables :

X : classical raw score of examinee (= sum of items) ;

P : classical index of difficulty of item (= proportion of examinees giving correct response) ;

θ, $\hat{\theta}$: examinee's ability generated or estimated from IRT procedures ;

$\hat{T}$: examinee's estimated "true score" computed from other IRT estimates ;

a, b, $\hat{a}$, $\hat{b}$: item's discrimination and difficulty index respectively, generated or estimated from IRT procedures.

For each of the required 30 iterations, each Monte Carlo produced the following statistics :

- test-retest reliability coefficient (measured with Pearson's r correlation coefficient on same variable at pre-test vs. post-test), for variables X, P, $\hat{\theta}$, $\hat{a}$, $\hat{b}$;
- Pearson's r correlation on different quantities at pre-test, for pairs of variables ($\hat{T}$, X), ($\hat{T}$, $\hat{\theta}$), (θ, $\hat{\theta}$), (θ, X), ($\hat{\theta}$, X), (a, $\hat{a}$), (b, $\hat{b}$), ($\hat{b}$, P);
- mean and standard deviation at pre-test for variables a, b, θ, $\hat{a}$, $\hat{b}$, $\hat{\theta}$, X;
- mean absolute error for $\hat{\theta}$, i.e. mean of $|\hat{\theta} - \theta|$;
- maximum absolute error for $\hat{\theta}$, i.e. max of $|\hat{\theta} - \theta|$.

Parametric data generated by our programs to simulate the random response protocols, i.e. the "true" θ, a and b values used for each Monte Carlo run, were tested for consistency with our sets of corresponding hyper-parameters (μ and σ^2) and proved unbiased and adequate (with no significant departure from due values).

Statistical analyses presented hereafter use either crossed ANOVA designs, Student's t tests or Pearson's correlation coefficients. Unless stated otherwise, statistical significance is at the 0.01 level or better. Finally, in order to give the reader some appreciation of effect size, Hays' (1981) omega squared (ω^2) index of experimental efficacy is occasionally produced ; the index is derived in the usual way from ANOVA's expectancy formula for mean square, under a fixed effects model.

Statistical analysis

In a first section, we present results on the test-retest reliabilities of $\hat{\theta}$ and X, "pertinence measures" for $\hat{\theta}$, the spread of $\hat{\theta}$ distributions and their accuracy in order to answer our main questions about the matching of the estimated abilities ($\hat{\theta}$) with their corresponding generated values (θ). We shall also examine the factors that influence the consistency, reliability and other statistical characteristics of $\hat{\theta}$, and the correspondence between $\hat{\theta}$ and the classical raw scores X across conditions. The second section will bear on item parameters' estimation, namely the test-retest reliability, pertinence and accuracy of $\hat{b}$ and $\hat{a}$ estimates.

Around the ability level estimate ($\hat{\theta}$)

Test-retest reliability of $\hat{\theta}$ and X

Globally, across conditions, the levels of test-retest reliability for $\hat{\theta}$ and raw score X are equivalent ($F < 1$). Data in Table 1 were taken from Case 1 under standard ($\mu_\theta = 0$, $\mu_b = 0$) conditions. Both reliabilities increase as a function of the number of items (k) ($F = 1579.27$ and 1756.36, $df = 3, 348$, $\omega^2 = 0.929$ and 0.936, for $\hat{\theta}$ and X respectively), and they follow almost exactly the Spearman-Brown prediction formula (used to predict the reliability of a test whose number of items has been changed [Lord & Novick, 1968; Bertrand & Blais, 2004]) , an expected result for the X score but one that comes somewhat as a surprise for the θ estimate. Furthermore, reliability of θ estimates increases with the number of examinees ($F = 12.85$, $df = 2, 348$, $\omega^2 = 0.062$) while this is not the case for scores X ($F < 1$). In fact, θ estimates tend to be stabilized especially when N increases from 100 to 500 subjects, which is not he case for X.

This last result is interesting, because one of fundamental assumptions of CTT is that the reliability of scores X depends on k, the number of parallel items, via the accumulation of true variance, but it does not depend on N. Now, in order to estimate examinees' θs, our IRT procedure first obtains estimates for item parameters a and b, estimates which appear to be stabilised by an increase of N, the number of protocols processed (see below). Thus, the observed gain in reliability of θ may be a corollary of the increase in item parameters' reliability.

Data from cases 2 and 3 of the experimental design confirm the above results. In case 2, raw score reliability was "imposed" at $\rho(X, X) \approx 0.40$ across protocols with diverse number of items (k) by varying hyper-parameter μ_a, and at $\rho(X, X) \approx 0.80$ in case 3. Both the observed test-retest

Table 1. Test-retest reliability of $\hat{\theta}$ and X as a function of k and N

	k				N		
	10	30	50	100	100	500	1000
$r(\hat{\theta}_1, \hat{\theta}_2)$	0.342	0.613	0.729	0.846	0.621	0.642	0.642
$r(X_1, X_2)$	0.345	0.615	0.721	0.842	0.626	0.635	0.633

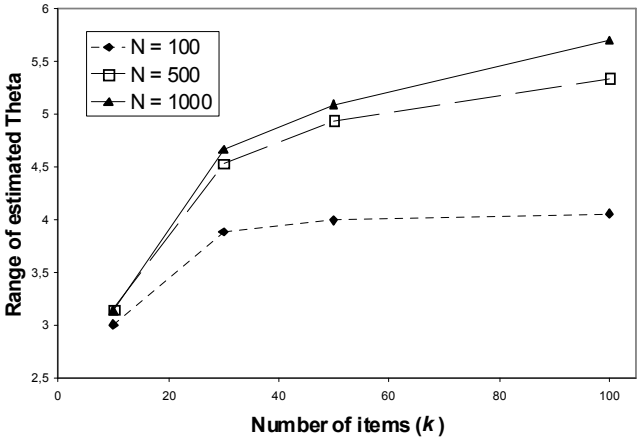

Figure 1. Range of estimated $\hat{\theta}$ distribution as a function of k and N (case 1)

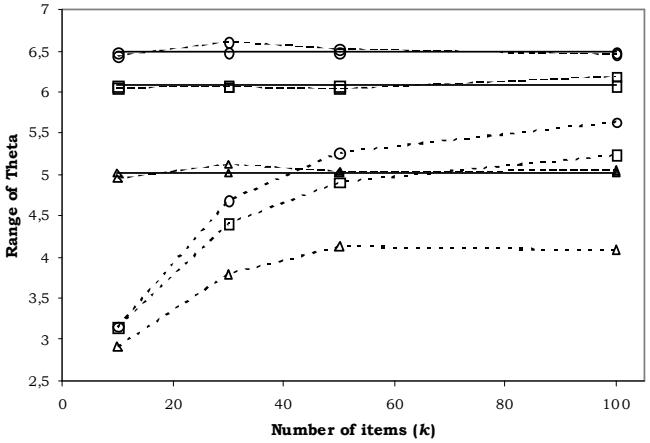

Figure 2. Range of theta values of N = 100 (△), 500 (□) and 1000 (○) for generated (broken line), estimated (dotted line) and asymptotic normal values (full line).

reliability values for X and $\hat{\theta}$ were close to target values, respectively with RMS indices of 0.015 and 0.029 in case 2 and 0.003 and 0.008 in case 3. The only noteworthy ANOVA results relate to the N × (X, $\hat{\theta}$) interaction, for all k confounded : $r(\hat{\theta}_1,\hat{\theta}_2)$ increases consistently from N = 100 to N = 1000 (F = 275.40 and 200.26 for cases 2 and 3, df = 1, 348), whereas $r(X_1, X_2)$ does not (F = 7.30 and 3.23).

Pertinence of measures for θ

In this Monte Carlo investigation, the true examinee's ability level was defined by his θ parameter, so that the pertinence of various estimating functions of ability can be directly assessed. Correlations of θ with diverse ability estimates ($\hat{\theta}$, X,$\hat{T}$) will now be scrutinised.

Correlations $r(\theta, X)$ and $r(\theta, \hat{\theta})$ behave similarly to corresponding reliability coefficients $r(X_1, X_2)$ and $r(\hat{\theta}_1, \hat{\theta}_2)$, except that they are stronger and vary more slowly as a function of k. In fact, the ratio between the two sets of indices amounts to the ratio between $r(T, X)$ and $r(X, X)$ in test theory, i.e. $r(T, X) = \{ r(X, X) \}^{½}$. These correlations also increase with k as the square root of the Spearman-Brown formula, and they similarly interact with N, the number of examinees. Averaging over all combinations of case 1, it seems worthwhile reporting that the means of $r(\theta, X)$ and $r(\theta,\hat{\theta})$ are quasi equal, i.e. 0.785 and 0.786 respectively (F < 1), the more so if we consider that quantities θ and $\hat{\theta}$ are generated / estimated on a standardised continuous scale about the normal probability model, and score X is a crude binomial-like count with its well-known ceiling and floor effects. The assumed specificity and advantage of θ and IRT scaling do not stand out here.

In order to enable us to delve into the inter-relations of θ, $\hat{\theta}$, X and $\hat{T}$, we ran a special Monte Carlo experiment with fixed conditions N = 500, k = 30, μ_θ = 0, μ_a = 0.5 and μ_b = 0. First, we obtain equivalent pertinence coefficients for $\hat{\theta}$ ($r(\theta, \hat{\theta})$ = 0.799) as for X ($r(\theta, X)$ = 0.792). Now, if we correlate each observed score-value (X') to the mean of all generated θ-values associated with it, we obtain $r(\overline{\theta}, X')$ = 0,980, a quasi perfect match. Also, the raw $r(\hat{\theta}, X)$ = 0.9746 jumps to $r(\hat{\theta}, X')$ = 0.9974 when we regroup equal X scores and average the concomitant $\hat{\theta}$ values. Finally, considering $r(\hat{\theta}_1, \hat{\theta}_2)$ = 0.633, the $r(\theta, \hat{\theta})$ corrected for attenuation becomes $0.799 / \sqrt{0.633} \approx 1.00$, similarly to $r(X_1, X_2)$ = 0.625 and corrected $r(\theta, X_1) = 0.792 / \sqrt{0.625} \approx 1.00$: this result seems to indicate that all the information (or portion of "true variance") contained in the true θ values is equally in the $\hat{\theta}$ and the X estimates, hence that these estimates are linearly equivalent.

We verified the correlation between $\hat{\theta}$ and X, our two estimates of subject's ability level. Minimum, maximum and mean values across 360 various iterations were r = 0.873, 0.994 and 0.974. Correlations grow as a function of k (F = 142.53, df = 3, 348, ω^2 = 0.541) and of N (F = 55.11, df = 2, 348, ω^2 = 0.231), with a slight interaction (F = 4.22, df = 6, 348, ω^2 = 0.051) reflecting the fact that the effect of N diminishes as k increases.

The preceding results suggest that the true variance available in the generated θ sample is transferred to the X score distribution as well as in the estimated $\hat{\theta}$. Moreover, considering the forceful increase of correlation from individual values, i.e. $r(\theta, X)$ = 0.792, to regrouped values, i.e. $r(\overline{\theta}, X')$ = 0,980, the supernumerary $\hat{\theta}$ values produced by the IRT estimation procedure appear to convey no more information, except some noise. Computations done with estimated true scores ($\hat{T}$) lend the same results and point to the same conclusions.

Spread of $\hat{\theta}$ distribution

Ability parameters (θ) in case 1, under conditions $\mu_\theta = \mu_b$

Table 2. Accuracy of $\hat{\theta}$ as a function of k and N

	k				*N*		
	10	30	50	100	100	500	1000
ave $\lvert\hat{\theta}-\theta\rvert$	0.648	0.493	0.415	0.321	0.485	0.462	0.461
max $\lvert\hat{\theta}-\theta\rvert$	2.545	1.945	1.673	1.281	1.697	1.878	2.008

= 0, were generated from a normal distribution having mean 0 and standard deviation 1, generated values being consistent with these characteristics, as mentioned earlier. As for *estimated* ability data ($\hat{\theta}$), their observed mean was consistently equal to 0, a result that is ascribable to the EAP estimation procedure and implementation. The spread of $\hat{\theta}$ values was assessed by two statistics, range and standard deviation; the two bearing similar results, we report only the former. Figure 1 depicts the evolution of the range as a function of *k* (F = 311.66, df = 3, 348, ω^2 = 0.721) and N (F = 692.53, df = 2, 348, ω^2 = 0.793), with an interaction effect (F = 35.31, df = 6, 348, ω^2 = 0.364), the increase of range getting slower as N goes from 1000 to 100.

In order to get a more thorough understanding of the above results and establish them firmly, we ran another series of Monte Carlo experiments, this time using 100 replications under standard conditions of case 1 (μ_θ = 0, μa = 0.5, μb = 0) and measuring the range of generated (θ) and estimated ($\hat{\theta}$) ability levels : results appear in Figure 2. Data for estimated $\hat{\theta}$ in Figure 2 (dotted lines) match closely those shown in Figure 1, re-enacting the varying influence of k on the spread. On the other hand, generated θ values seem to be unperturbed by k, and, for each level of N, they agree satisfyingly with their expected values under the normal distribution[5]. Since the spread of $\hat{\theta}$ increases as a function of k but only up to a certain maximum depending on N, two questions must be tackled : What mechanism can be invoked to explain the increase on k, and what blocks its effect at each N ?

Accuracy of $\hat{\theta}$ estimates

We now look into the accuracy of the ability estimate $\hat{\theta}$ by examining different measures of the distance between the $\hat{\theta}$ value and its target θ. Data in Table 2 stem from case 1 under standard conditions μ_θ = 0 and μ_b = 0 : one is the average of the absolute deviations between $\hat{\theta}$ and θ across the N pseudo respondents, the other is the maximum absolute deviation. All variations are significant at more than 0.01 alpha level. *Ave* $\lvert\hat{\theta}-\theta\rvert$ decreases more rapidly as a function of *k* (ω^2= 0.937) than of N (ω^2= 0.115) ; note that quantity *ave* $\lvert\hat{\theta}-\theta\rvert$ follows closely

$$\sqrt{\tfrac{2}{\pi}\left(1-r_{\hat{\theta},\hat{\theta}}\right)},$$

which is the expected absolute deviation for a normal variate with standard deviation given by the standard error of $\hat{\theta}$. For max $\lvert\hat{\theta}-\theta\rvert$, it decreases with *k* (ω^2= 0.785) but increases with N (ω^2= 0.216), as expected[6]. Considering that the true (i.e. generated) θ distribution has a standard deviation of 1, the reported differences between true and estimated θs appear somewhat large, i.e. near to 0.5 for the *Ave* measure and to 2.0 for the Max measure.

Now, what does happen if we leave behind us the reassuring environment of « standard conditions » and explore new parametric conditions with various μ_θ and μ_b? For the sake of simplicity, the following analyses were limited to sub-conditions N = 500 and *k* = 30, and they are entirely representative of all our N and *k* combinations.

Table 3, below, reports the means of $\hat{\theta}$, X and $\hat{b}$ in situations for which μ_b = 0 and μ_θ = 0, 1 and 2 respectively. Note that, here as everywhere, the mean values of the N estimated $\hat{\theta}$ are consistently 0, a likely corollary of the two-phase and EAP estimation procedure put to work. Raw scores increase with μ_θ (F = 295.67, df = 2, 87, ω^2 = 0.868), as expected. The $\hat{b}$ estimates vary also (F = 696.86, df = 2, 87, ω^2 = 0.939) but they do so contrary-wise and, in view of their values, compensate almost exactly the variation of μ_θ.

Table 3. Means of parameter estimates under conditions N = 500, k = 30, μ_b = 0 as a function of μ_θ (30 replications)

mean	μ_θ = 0	μ_θ = 1	μ_θ = 2
$\hat{\theta}$	0.000	0.000	0.000
X	14.99	18.07	21.27
$\hat{b}$	0.012	–0.950	–2.095

The means in Table 4 represent situations with fixed μ_θ = 0 and with varying μ_b (= 0, 1, 2). Here again, next to $\hat{\theta}$ = 0, which is expected though still astonishing[7], we observe a decrease in *X* scores (F = 688.26, df = 2, 87, ω^2 = 0.939) as item difficulty (μ_b) increases, which is also reflected in the means of $\hat{b}$ (F = 757.60, df = 2, 87, ω^2 = 0.944). Here is thus a quasi plausible parametric outcome, stemming from a standard, centered population (with μ_θ = 0).

Finally, Table 5 renders three situations wherein hyper-

Table 4. Means of parameter estimates under conditions N = 500, k = 30, μ_θ = 0 as a function of μ_b (30 replications)

mean	μ_b = 0	μ_b = 1	μ_b = 2
$\hat{\theta}$	0.000	0.000	0.000
X	14.99	11.53	8.85
$\hat{b}$	0.012	1.073	1.989

Table 5. Means of parameter estimates under conditions N = 500, k = 30, and $\mu_\theta - \mu_b = 0$ (30 replications)

mean	$\mu_\theta = 0, \mu_b = 0$	$\mu_\theta = 1, \mu_b = 1$	$\mu_\theta = 2, \mu_b = 2$
$\hat{\theta}$	0.000	0.000	0.000
X	14.99	15.00	15.01
$\hat{b}$	0.012	–0.029	0.006

parameters $\mu_\theta = \mu_b$ are compared. Here, the interaction between the θ and b parameters occurs at generation time, i.e. the relocation of the distributions of θ and b is annihilated by subtraction in the exponent part of equation (1), "$(\theta_r - b_i)$", so that all our observed means are at center. This consequence, even though it is evident, pinpoints the essential and mutual indeterminacy of the θ and b scales, which affects all IRT models, from the 1-parameter model upward.

Cases 4a-4d of our experimental design make use of witness pseudo respondents, i.e. respondents for whom the same pre-test protocol is used at post-test and whose ability level is then estimated among new companion protocols. Conditions for estimation were N = 500 respondents (among which 10 were retained as witnesses), k = 30 items, item discrimination and difficulty hyper-parameters $\mu_a = 0.5$ and $\mu_b = 0$. As usual, 30 Monte Carlo replications were performed; for each replication, 10 estimation samples were taken in order to accumulate 100 witnesses for purpose of statistical analyses.

Test-retest reliability of witness data are reported in Table 6. Data from Case 4a were used to validate our sampling scheme, and they mimic the corresponding data from Case 1. As can be seen, the reliability coefficients *per se* are but slightly affected by their new estimation environment, whether 98% (= 490 / 500) new protocols emanate from the same original companions (Case 4b), from new companions having the same ability level (Case 4c) or even from new and much more talented companions (Case 4d). For all cases, the two $\hat{\theta}$ values estimated from the same protocol are highly correlated. The near-to-perfect reliability coefficient means that individual $\hat{\theta}$s keep a linear relation one to the other, in the guise of $\hat{\theta} = b_1 \hat{\theta}_1 + b_0$: as we shall see, the accuracy problem resides in the "b_0" component.

To throw some light on the intricacies of our problem, we ran yet another series of estimation runs for Case 4d, again with 30 replications, thereby collecting new types measurement. The salient results follow.

Firstly, the b parameter obtains a $\overline{\hat{b}_1}$ = 0.000 at pre-test and $\overline{\hat{b}_2}$ = –0.967 at post-test (t = 25.08 to 36.86[8], df = 299). This negative shift of the difficulty parameter estimates (notwithstanding the constant $\mu_b = 0$) reflects the positive shift of the population level (from $\mu_\theta = 0$ to $\mu_\theta = 1$), the two-phase estimation procedure banking on a 0-centered normal population.

Secondly, the $\hat{\theta}$ estimates glide from $\hat{\theta}_1$ = –0.013 to $\hat{\theta}_2$ = -0.629 (t = 30.83 to 43.57, df = 99), a negative shift ascribable to the mixing of our witnesses (coming from a $\mu_\theta = 0$ population) with brighter companions (coming from a $\mu_\theta = 1$

Table 6. Test – retest reliability coefficients for 4 indices in Case 4

	Description	$r(X_1,X_2)$	$r(\hat{\theta}_1,\hat{\theta}_2)$	$r(\hat{T}_1,\hat{T}_2)$	$r(PR_1,PR_2)$†
Case 4*a*	Same respondents ($\mu_\theta = 0$) at post-test, new protocols for all at post-test	0.621	0.623	0.640	0.606
Case 4*b*	Same respondents ($\mu_\theta = 0$) at post-test, 10 witness protocols + 490 new protocols at post-test	-*	0.981	0.998	0.962
Case 4*c*	New respondents ($\mu_\theta = 0$) at post-test, 10 witness protocols + 490 new protocols at post-test	-*	0.980	0.982	0.996
Case 4*d*	New respondents ($\mu_\theta = 1$) at post-test, 10 witness protocols + 490 new protocols at post-test	-*	0.978	0.981	0.954

† PR = percentile rank * $r(X_1,X_2) = 1$ by definition for witness protocols.

Table 7. Reliability and pertinence of difficulty indices $\hat{b}$ and P

	N = 100	N = 500	N = 1000		N = 100	N = 500	N = 1000
$r(\hat{b}_1, \hat{b}_2)$	0,731	0,897	0,946	$r(b, \hat{b}_1)$	0,848	0,949	0,972
$r(P_1,P_2)$	0,850	0,962	0,983	$r(b, P_1)$	–0,897	–0,949	–0,964

Table 8. Accuracy of $\hat{b}$ as a function of k and N

	k				*N*		
	10	30	50	100	100	500	1000
ave $\lvert\hat{b} - b\rvert$	0.332	0.283	0.280	0.268	0.456	0.246	0.171

population). The mechanism of this negative relocation of our witnesses is as follows. The better part of post-test protocols comes from a high-level ($\mu_\theta = 1$) *true population* : in order to maintain a centered ($\hat{\theta}_1 = 0$) *estimated population*, the estimation procedure is forced to assume easier items, thus producing lower *b* indices. Now, lower *b* indices, and easier items, applied to unchanged witness protocols entail a concomitant reduction in the estimated ability levels. Equivalently, the witness protocols at post-test are processed amidst higher-graded protocols, the set of which is to be matched with a 0-centered population. Consequently, the small batch of our 10 witnesses is thus downgraded along the ability axis. A phenomenon analogous to the above occurs for the percentile rank (PR) of witnesses, which we computed. At pre-test, we obtain $\overline{\mathrm{PR}}_1 = 49.42$, close to 50 as expected, and $\overline{\mathrm{PR}}_2 = 28.79$ at post-test ($t = 20.18$ to 27.74, $df = 99$), a negative shift imposed by the fact that ranks are computed for all 500 respondents among which 490 have now enhanced response protocols.

Thirdly, the true score estimates ($\hat{T}$), computed at each time from $\hat{\theta}$, $\hat{a}$ and $\hat{b}$ estimates, change from $\hat{T}_1 = 14.95$ at pre-test (near to the $\frac{1}{2}k = 15$ target) to $\hat{T}_2 = 16.03$ at post-test ($t = -10.21$ to -9.71, $df = 99$), a small but consistent and highly significant *positive* shift. Though paradoxical, this result may tentatively be explained by a differential effect of the change in μ_θ from 0 to 1. On the one hand, because our 10 (out of 500) witnesses originate from all strata of the $\mu_\theta = 0$ population, some may be better gifted and compatible with $\mu_\theta = 1$ subjects, and consequently their mean estimated ability decreases from about 0 ($\overline{\hat{\theta}}_1 = -0.013$) to only $\overline{\hat{\theta}}_2 = -0.629$ instead of –1. On the other hand, because the procedure for item-parameter estimation is confronted with 98% high-grade $\mu_\theta = 1$ protocols, it forces the $\hat{b}$ estimates at post-test down to $\overline{b_2} = -0.967$, quite near to –1. Hence, even if subjects' ability levels have been lowered at post-test, they were administered items of an even lower difficulty level, resulting in a small rise of their predicted $\hat{T}$ value.

Lastly, again for the accuracy of $\hat{\theta}$ estimates, we calculated the mean absolute difference (MAD) between true θ and estimated $\hat{\theta}$ at the two estimation times. At pre-test, we observed $\mathrm{MAD}_1 = 0.492$ (range 0.414 to 0.605), and at post-test, $\mathrm{MAD}_2 = 1.090$ (range 0.870 to 1.348), a significant increase ($t = -27.12$, $df = 32$), betraying an important loss of accuracy for our witness $\hat{\theta}$.

Around the item parameters estimates

Reliability, pertinence and accuracy of $\hat{b}$ estimates

The classical item difficulty estimate *P*, which (oddly) designates the proportion of correct responses for the item, is the analogue of IRT's $\hat{b}$ parameter, and their statistical behaviours can be securely compared. Such comparisons are effected in Table 7 : data originate from Case 1, with $\mu_\theta = 0$, $\mu_b = 0$ and $\mu_a = 0.5$.

Item parameters *a*, *b* and *P* being estimated across respondents, it is to be expected that both reliability and pertinence coefficients benefit from an increase of their number N. Reliability of *P* is globally somewhat higher than that of $\hat{b}$ ($F = 738.69$, $df = 1, 348$, $\omega^2 = 0.672$), the difference diminishing as N increases ($F = 79.07$, $df = 2, 348$). For pertinence coefficients (right of Table 7), the initial advantage of *P* at N = 100 ($F = 631.81$, $df = 1, 348$) vanishes at N = 500 ($F < 1$) and turns upside down at N = 1000 ($F = 17.55$).

As for the accuracy of $\hat{b}$, we must recall first that, in the realm of our IRT procedures and programs, hyper-parameters μ_b and μ_θ play a compensatory game by virtue of which the resultant $\hat{\theta}$, the mean of the estimated θ distribution, is put equal to 0. This means, for instance, that a $\mu_\theta = \delta$ distribution of true abilities will engender a $\mu(\hat{\theta}) = 0$ distribution of estimated abilities and a concomitant $\mu(\hat{b}) = -\delta$ distribution of difficulty indices. Given this caveat, the means $\hat{b}$ are unbiased ($\approx -\delta$), with no variance effects of N or of *k*. As shown in Table 8, the relative accuracy of $\hat{b}$ increases with *k* ($F = 14.38$, $df = 3, 348$, $\omega^2 = 0.100$) and more so with N ($F = 520.54$, $df = 2, 348$, $\omega^2 = 0.743$), these effects diminishing somewhat ($F = 2.48$, $df = 6, 348$, $p < 0.05$, $\omega^2 = 0.024$) toward high values of N and *k*.

Reliability, pertinence and accuracy of $\hat{a}$ estimates

The global behaviour of the discrimination parameter estimate ($\hat{a}$) has been studied through conditions of cases 2 and 3, which offered 8 different values of hyper-parameter μ_a (from 0.165 to 1.843). The bias of $\hat{a}$ (measured with $\hat{a} - \mu_a$) is generally positive; in contrast to generated *a*'s, where $\overline{a} = \mu_a$ ($R^2 = 0.9805$), we observe $\hat{a} = 0.1079 + 0.9384 \times \mu_a$ ($R^2 = 0.9355$). Positive bias is higher for small μ_a's, which correspond to high values of *k* in our design ($F = 24.70$, $df = 11, 1044$), and it decreases steadily toward 0 as N increases ($F = 214.38$, $df = 2, 1044$). Computations with ratios $\hat{a} / \mu_a$

Table 9. Reliability, pertinence and accuracy of â as a function of k and N

	k				*N*		
	10	30	50	100	100	500	1000
$r(\hat{a}_1, \hat{a}_2)$	0.282	0.508	0.544	0.571	0.198	0.538	0.693
$r(a, \hat{a}_1)$	0.524	0.684	0.719	0.742	0.429	0.736	0.837
ave $\lvert\hat{a} - a\rvert$	.174	0.132	0.126	0.127	0.237	0.109	0.073

render similar effects.

As for individual $\hat{a}$ estimates, Table 9 documents some of their properties : data come from case 1, in the standard ($\mu_\theta = 0$, $\mu_b = 0$) conditions. Reliability increases with k ($F = 40.87$, $df = 3, 348$, $\omega^2 = 0.249$) and more so with N ($F = 201.16$, $df = 2, 348$, $\omega^2 = 0.527$), without interaction effects ($F = 1.11$, $df = 6, 348$). Pertinence values follow a similar pattern, increasing with k ($F = 34.07$, $\omega^2 = 0.216$) and N ($F = 212.38$, $\omega^2 = 0.540$), without interaction ($F = 1.14$). Accuracy, measured by *ave* $\lvert\hat{a} - a\rvert$, gets better as a function of k ($F = 49.24$, $\omega^2 = 0.287$) and of N ($F = 898.98$, $\omega^2 = 0.833$) ; a small interaction term ($F = 4.09$, $df = 6, 348$, $\omega^2 = 0.049$) indicates that effect of k diminishes as N increases.

Among many others, the preferred index of "item discrimination" in CTT is probably the item-test correlation coefficient, $r(y_j, X)$, where y_j is the subject's 0/1 response at item j, and X is his number of good responses across k items (where $X = \Sigma\ y_j$). We correlated this index with the $\hat{a}_j$ parameter estimate : Table 10 presents the correlations obtained.

First, the individual correlations (per iteration) range from 0.779 up to 0.997 and have a global average of 0.976. Generally they increase with k ($F = 44.76$, $df = 3, 348$, $\omega^2 = 0.267$) and N ($F = 67.07$, $df = 2, 348$, $\omega^2 = 0.269$). A curious interaction occurs ($F = 11.48$, $df = 6, 348$, $\omega^2 = 0.149$), the correlations for N = 100 curving down unexpectedly at $k = 100$ (a fact which explains the drop of values at $k = 100$, in Table 10).

3. DISCUSSION

About the ability parameter estimate

Interesting features of Monte Carlo studies such as this one include the cheap abundance of generated data and the fact that data models are explicitly defined and implemented. In the case of the present study, true ability and item parameters were put on stage, together with their statistical estimates, which derived from response protocols obtained through an explicit 2-parameter logistic IRT model. Thus, the IRT model's properties were assured *per definition*, and conclusions hereof can be safely drawn.

Regarding the invariance of θ ability estimate in IRT, our data plainly demonstrate two things. Firstly, as Table 3 shows, the $\hat{\theta}_i$ estimate is generally biased, bias coinciding accidentally with zero when the associate population location parameter (μ_θ) is zero. This bias pertains to the indeterminacy of the θ and b scales, in fact to the indeterminacy of the difference $(\theta - b)$ in the definition of the IRT model (see formula 1 at page 3). The currently applied two-phase estimation procedure, used also in this study, settles the indeterminacy by anchoring the $\hat{\theta}$ estimates in a finite distribution with mean 0, i.e. $\mu(\hat{\theta}) = 0$, while allowing a shift of the $\hat{b}$ distribution to accommodate examinees' response patterns. Another factor infringing invariance is k, the number of items in the estimation set, the spread (range and standard deviation) of estimated $\hat{\theta}_j$ correlating with k (see Fig. 1 and 2). Secondly, observed values of correlations between true θ_i and estimated $\hat{\theta}_i$, our so-called pertinence coefficients, give a paradoxical support to the "congruence" aspect of invariance. It is true to say that estimates $\hat{\theta}_i$ are linearly related to their parametric counterparts θ_i, the paradox being that (1) the correlation level between the two obeys standard theorems of CTT, theorems whose demonstration is based on the interplay of true and error variance, and (2) expectedly enough, the same correlation levels are observed between θ_i and raw score X_i. The threshold of "$r \geq 0.90$" for declaring invariance appears quite irrelevant in this context. Moreover, data from our witness protocols show patently that the level (or bias) of the $\hat{\theta}_i$ estimate is adjusted to the context of the companion respondents with which it is processed – notwithstanding the fact that the adjusted $\hat{\theta}_i$ is based upon an unchanged

Table 10

Mean correlation (Pearson's r) between $\hat{a}_j$ and $r(y_j, X)$

	k				N		
	10	30	50	100	100	500	1000
$r\{\hat{a}_j, r(y_j, X)\}$	0.958	0.983	0.986	0.977	0.961	0.983	0.984

response protocol.

Summarising the above discussion and the detailed evidence in our data, we assert that the concept of "invariance", given as a distinctive asset of item response theory, is over-defined and overrated. The $\hat{\theta}_i$ estimates are not invariant across a shift in the population location parameter (μ_θ) and their spread and positioning are influenced by the number of items. If $\hat{\theta}_i$ estimates appear to be invariant across sets of items, it is because the $\hat{\theta}_i$ distribution is formed anew using $\mu(\hat{\theta}) = 0$, notwithstanding the actual difficulty levels (μ_b) of items. Thus, the alleged "invariance of the IRT ability estimate" should be changed to "linear equivalence, tainted with bias indeterminacy". Moreover, the observed linear, or "congruence", properties of $\hat{\theta}_i$ are entirely shared by classical score *X*, apart from the fact that only *X* scores levels vary coherently with true θ levels. Hence, in our opinion, "invariance" does not hold for the $\hat{\theta}$ ability estimate and, above all, its remaining "congruence" (or linear) properties do not constitute a distinctive asset, as they characterize also the classical *X* score measure. Finally, the apparent superiority of $\hat{\theta}_i$ estimates in terms of discriminating capacity[9] is not corroborated by a superior reliability level or by a better efficiency to discriminate respondents, as was shown earlier.

About the item parameters' estimates

Accessorily, the present Monte Carlo rendered some information on the properties of item parameter estimates, the $\hat{a}$ et $\hat{b}$ of IRT as well as $r(y_j, X)$ and *P* in CTT. Due to the (θ, *b*) indeterminacy mentioned above, the $\hat{b}$ estimate is intrinsically biased, its location parameter $\mu(\hat{b})$ playing a compensatory role with $\mu(\hat{\theta})$ in fixing $\mu(\hat{\theta}) = 0$ in phase 1 of the two-phase estimation procedure. Apart from that, $\hat{b}$ and *P* are linearly congruent with the true *b* parameter (see Table 7), the classical *P* index being somewhat more reliable. No crude bias effect was observed for the $\hat{a}$ estimate, were it not for a slight positive bias depending upon hyper-parameter μ_a. Our data contain no information specific to the "invariance" of the $\hat{a}$ estimate, as the experimental design of the study included only shifts in location (via hyper-parameters μ_θ and μ_b) but not in range (all θ and *b* distributions were controlled with $\sigma^2 = 1$). Estimated $\hat{a}$'s were not perturbed by shifts of location in ability or difficulty parameters, and they correlated nicely with true *a* values (see Table 9).

4. CONCLUSION

Item response theory's estimates of ability ($\hat{\theta}$) are not invariant across a change of the estimation context, be it a shift in the ability level of co-examinees or in the global difficulty level of items. Wells et al. (2002) have already documented such variant effects under small changes in subgroups of items, but the real snag comes from the intrinsic indeterminacy of the (θ, *b*) pair in all IRT models, e.g. $P_j(\theta_r) = [\ 1 + \exp(\ -a_j(\theta_r - b_j)\ ]^{-1}$, the operating characteristic of $P_j(\theta_r)$ being the difference $(\theta_r - b_j)$ where each shift of θ can be compensated by an equal shift of *b*. IRT estimation procedures, like the two-phase procedure employed here, cannot overcome this indeterminacy, which results in the fact that $\hat{\theta}$ distributions are generally biased and arbitrarily centered on $\mu(\hat{\theta}) = 0$.

On the other hand, the $\hat{\theta}$ estimate displays nice linear, or "congruence" (Hambleton et al., 1991) properties : its reliability levels are comparable to those of classical raw score *X*, and its pertinence, i.e. correlation of the estimate ($\hat{\theta}_i$) with the true individual parameter (θ_i), is also quite good, in fact, it is comparable again to the correlation between θ_i and raw score X_i ! These linear properties are influenced by the size of their estimation basis (the number of items) and statistically consistent ; indeed, they seem to reflect the proportional amount of true variance, a standard result in CTT and one that does not appear to be grounded in the algebraic framework of item response theory. As for the relative advantages of item parameter estimates of IRT versus CTT, IRT's estimate of item difficulty $\hat{b}$ correlated highly with classical *P* difficulty index, index *P* being somewhat more reliable and better correlated with the true parametric *b* value. The two discrimination indices also compared well, IRT's $\hat{a}$ coefficient being a little more reliable and better linked to the *a* parameter than classical item-test correlation $r(y,X)$.

It may be pertinent here to restate that the generic IRT model is, up to now, the only conceptual apparatus that can pretend to be a true "model" of what goes on between the respondent and the set of items that confronts him. It is a first-order (no interactions nor sequential processes are assumed), stimulus-response model, but, even then, it goes a long way beyond the crude axiomatic basis of CTT. When we turn to concrete applications, though, this good model turns up flawed with an intrinsic indeterminacy and with grave estimation problems. Consequently, when on the practical side, the hoped-for theoretical merits of IRT estimates become tarnished and should be judged on a psychometric basis at a par with the classical estimates *X*, *P* and $r(y, X)$, who behave as well if not better (Fan 1998 ; Frenette et al. 2007).

Even if the procedural and parametric settings of this simulation study matched those of current IRT applications, their limitations are present and must be overcome. How do IRT estimates fare under estimation procedures different from the ones employed here and, above all, what are the essential changes entailed by a one-phase, joint estimation

procedure (for both ability and item parameters) ? How do test-retest reliability levels of ability estimates ($\hat{\theta}$, X) compare, when item parameters are estimated only once, at test time? And, finally, are the concepts of "invariance" and "dimensionality" really cardinal in the epistemological definition of a response model (Blais, 1987 ; Frenette et al., 2007 ; Wells et al., 2002), and could they not be replaced by the more universal descriptors used in statistical estimation theory ? The creative potential of the generic IRT model has not been exhausted by this or all other studies, and much has yet to be harvested with its aid.

References

Baker, F.B. & Kim, S.-H. (2004). *Item response theory : Parameter estimation techniques* (2nd ed.). New York, Marcel Dekker.

Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items*. Québec : Presses de l'Université du Québec.

Bock, R.D. & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*, 431-444.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *39*, 1-38.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Frenette, E., Bertrand, R., Valois, P., Dussault, M. & Hébert, M.-H. (2007). Comparaison empirique des paramètres d'items/sujets de la théorie des réponses aux items et de la théorie classique des tests. *Publication des actes du 19ème colloque de l'admée-Europe, 1*(1), 1-14.

Freund, J.E. (1992). *Mathematical statistics* (5th ed.). Englewood Cliffs (NJ): Prentice-Hall.

Germain, S., Valois, P. & Abdous, B. (2008). Libirt: Item Response Theory Library (Version 0.8.4) [Computer software]. Available from http://libirt.sf.net/

Gulliksen, H. (1950). *Theory of mental tests*. New York : Wiley.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park (CA) : Sage.

Hays, W.L. (1981). *Statistics for the social sciences* (3rd ed.). New York : Holt, Rinehart and Winston.

Kendall, M. & Stuart, A. (1977). *The advanced theory of statistics. Volume 1 : Distribution theory* (4e édition). New York : Macmillan.

Laveault, D. & Grégoire, J. (2002). *Introduction aux théories des tests en sciences humaines* (2e édition). Paris : De Boeck.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading (Ma) : Addison-Wesley.

McKinley, R.L. (1989). Methods, plainly speaking: An introduction to Item Response Theory. *Measurement and Evaluation in Counseling and Development*, *22*, 37-57.

Owen, D.B. (1962). *Handbook of statistical tables*. Reading (MA) : Addison-Wesley.

Rup, A.A. & Zumbo, B.D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Si, C.-F. & Schumacker, R. (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing, 4(*2),137-181.

Sodoke, K., Raîche, G. & Nkambou, R. (2007). The adaptive and intelligent testing framework : PesonFit. Advanced Learning Technologies, 2007 ICALT, Seventh IEEE Conference on (2007), p. 715-717.

Stocking, M.L. &Lord, F.M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen D., Chen W.-H. & Bock R.D. (2003). *Multilog (version 7) [Computer software]* Lincolnwood, IL: Scientific Software International.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Educational Statistics*, *12*, 339-368.

Wells, C.S. Subkoviak, M.J. & Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(2), 77-87.

Wingersky, M.S., Barton, M.A. & Lord, F.M. (1982). *Logist user's guide*. Princeton, N. J. : Educational Testing Service.

Zimowski, M., Muraki, E., Mislevy, R.J. & Bock, R.D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models.* Chicago : Scientific Software.

Manuscript received May 18th, 2010
Manuscript accepted December 7th, 2010

Footnotes follow

[1] Expressed otherwise, "… the congruence between the two sets of estimates of each item parameter can be taken as an indication of the degree to which invariance holds" (Hambleton et al., 1991, p. 24).

[2] The functional relation $\rho_{XX} = f(k, \mu_a)$ was obtained empirically. On the one hand, we have $\rho_k \sim k\,\rho_1 / (1 + (k-1)\times\rho_1)$, as by the Spearman-Brown theorem and formula (see text) for scores based on k items. On the other hand, we found the approximate relation $\rho_1 \approx (0.2743\times\mu_a) / (0.2743\times\mu_a + 1)$, with $R^2 \approx 0.996$. Note that the (k, μ_a, ρ_{XX}) combinations were individually and precisely adjusted through Monte Carlo simulations.

[3] This computer freeware is available from http://libirt.sf.net/

[4] For instance, using real data from N = 400 respondents on k = 85 items (with dichotomous responses), the two-parameter a, b and θ estimates from BILOG-MG 3 correlated with EIRT's as 0.9982, 0.9876 and 0.9861 respectively. Artificial data on k = 10 items were generated by PersonFit (Sodoke, Raîche & Nkambou, 2007) for N = 100 "respondents" with ability levels (θ) uniformly spread from –3 to 3. Correlations between BILOG's and EIRT's estimates were 0.984, 0.999 and 0.998 for a, b and θ, and the maximum and mean absolute difference for θ were 0.257 and 0.039. Nearer to our setting, we generated data for N = 400 "respondents" with ability levels (θ) distributed as standard normal variates and k = 10 items (with uniform parameter distributions), and obtained correlations 0.9996, 0.9996 and 0.9997 for a, b and θ estimates, and maximum and mean absolute differences of 0.069 and 0.018 respectively, our confidence in the concordance of EIRT and BILOG-MG procedures being secured.

[5] The expected value of the range of N normal deviates is readily computed from expressions in Kendall and Stuart (1977, eq. 14.82, p. 362) or Owen (1962, p. 140). Note that, for the standard normal distribution, E(R) = 5.015, 6.073 and 6.483 for N = 100, 500 and 1000 respectively.

[6] The approximation $E\{ \max [\, |z_1|, |z_2|, \ldots, |z_N| \,] \} \times \sqrt{1 - r_{\hat{\theta},\hat{\theta}}}$, is well correlated to the max $|\hat{\theta} - \theta|$ measure, although with important haphazard discrepancies.

[7] Astonishing because the value recorded is a true zero, not a statistical (i.e. approximate) zero, and its variance is null across replications.

[8] Ranges (min *vs.* max) of Student's paired t values are given for the 30 replications.

[9] Except for identical response patterns, the expected number of different $\hat{\theta}_i$ estimates obtained is $min(N, 2^k)$, whereas the maximum possible number of different X scores is $k + 1$, a largely inferior value.