Some features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules [1]

Jean-Guy Blais

University of Montreal

Gilles Raîche

University of Moncton

[1] Paper presented at the 11th International Objective Measurement Workshop, New Orleans, april 2002.

1. **Introduction**

During an important part of the 20th century, the paper-pencil approach to tests has been privileged to gather certain pieces of information from individuals, subjects, candidates or exminees. When all the items of a test are the same for all examinees, according to the content as well as the number of items, the test can be label as fixed and invariable. When the number of items varies and when they can differ for each examinee, the test fits into the category of adaptive tests. In other words, it goes to the category of tests that are tailor-made owing to the problems met by the examinee at each item that is presented. The sequence of items then depends on the responses given to each item previously presented. The development of computers during the last twenty years has enabled a fast evolution of adaptive tests. In fact, while the models and branching strategies for adaptive testing have existed almost for fifty years, it is only recently that this testing strategy has been able to fully exercise its advantages over more conventional ones.

One of the most important characteristics of adaptive testing is the fact that it allows the administration of items with difficulty level corresponding to the proficiency level of the examinee. Contrary to the fixed and invariable paper-pencil tests, where all the items of the test are administered without considering the person's proficiency level, adaptive testing enables the administration of tailor-made tests, in such a way that the difficulty level of the items be neither too high or too low for the examinee. Hambleton, Swaminathan and Rogers (1991, p.145) underline that the number of items administered, like the duration of the administration, are then reduced comparing to a paper-pencil version of the test, without reducing the precision of the proficiency level estimate. According to Lord (1980, p.201), adaptive testing should provide a more precise estimate of the proficiency level, more specifically when the performance level is low or high.

Several characteristics of adaptive testing have received particular attention and been the object of some research work. Some authors have made comparisons between the proficiency level estimate obtained from different item response models (Dodd, de Ayala and Koch, 1995) and from different estimation methods (Chen, Hou and Dodd, 1998). Others have studied the influence of the dimensionality of the bank of items (de Ayala, 1992), the conformity to the postulate of local independence (Mislevy and Chang, 2000), the characteristics of the items' bank and the impact of different stopping rules on the proficiency level estimate (Dood, Koch and de Ayala, 1993). Comparisons have been made between item's selection rules (Chang and Ying, 1999), between methods used to assess the differential item functioning (Zwick, 1997) and between indices of adjustment of the proficiency level estimate or *person fit* (van Krimpen-Stoop and Meijer, 1999). But several aspects of the computer-based adaptive testing still remain to be studied.

Among the aspects of adaptive testing remaining to be studied, we have chosen to examine some characteristics of the statistics associated with the sampling distribution of the proficiency level estimate when the Rasch model is used. These characteristics enable to judge the meaning to be given to the proficiency level estimate obtained in adaptive testing and, as a consequence, can serve to illustrate the meaningfulness of the proficiency level estimate. If the sampling distribution of the proficiency level estimate follows a normal probability distribution, the precision of that estimate, when it is measured by its standard error, enables to determine a confidence interval for the proficiency level estimate. The determination of this interval is valid only when the sample distribution of the proficiency level estimate is symmetrical and mesokurtical, say neither too high nor too low.

According to Dodd *et al*. (1993), the characteristics of the sampling distribution of the proficiency level estimate in adaptive testing are affected by the use of different stopping rules. The aim of this research is precisely to study the effect of two stopping rules of frequent use on the characteristics of various statistics associated with the sampling distribution of the proficiency level estimate in adaptive testing. More exactly, it attempts to verify the impact of a stopping rule according to the *a priori* determination of the standard error of the proficiency level estimate and of a stopping rule according to the number of items administered on the second, third and fourth centered moments of the sampling distribution of the proficiency level estimate with the Rasch model.

2. Methodology

To exercise a strict control of the testing situation and needing that an important number of observations be available, we used a computer-based simulation. According to Harwell, Stone, Hsu and Kirisci (1996), a computer simulation with item response models is appropriate when we want to study the sampling distribution of the estimates or when we are comparing several methods oriented toward the same objective, without any possibility to obtain an exact analytical solution. According to the same authors, the utilization of a simulation also enables to manipulate more easily different factors at the same time, which is not always realizable in real conditions. The observations, viewed here as different values of the proficiency level, are generated at random, which characterizes a simulation of the Monte Carlo or stochastic type.

The initial values on the proficiency scale are randomly selected from a normal probability distribution with mean zero and variance one. A sample of 2000 values is generated in a random manner. It seems an appropriate sample size considering that sample sizes used in similar studies vary between 100 and 10,000, with a median value of 500. Adaptive tests are simulated for each of the 2000 random values of the proficiency level. Each simulation is performed according to a method for generating items response frequently used within the researches on item response models (Nicewander and Thomasson, 1999). For each value of the proficiency level generated, one obtains the response to each of the items by calculating the probability of obtaining a correct response to the item with the Rasch model as well as the value of the proficiency level. That probability is then compared to a random number, ranging from 0 to 1, drawn from a uniform probability distribution U(0,1). If the probability of obtaining a correct response to the item is superior to the drawn random number, the response to the item takes on the value 1, otherwise the response to the item takes on the value 0.

At all the simulated proficiency levels, the test begins with the administration of an item with a difficulty level, b_1, equal to 0, say the mean of the a priori distribution. The use of a constant starting difficulty level enables to assure that the proficiency level estimate obtained does not vary in relation to the difficulty level of the first item administered. We use Urry's method (Thissen and Mislevy, 1990, p.111; Urry, 1970, p.82) to select the next item. According to that method, the next item (j+1)th to be administered corresponds to an item with a difficulty level equal to the provisory proficiency level estimate obtained after the administration of the jth item. With the Rasch model, this selection rule allows to obtain the next item that provides maximum information; it is therefore equivalent to a strategy of information maximization. Moreover, it enables to choose the values of the difficulty parameter according uniquely to the selection rule in such a way that the characteristics of the items' bank do not affect the values of the proficiency level estimate.

The estimate of the proficiency level is obtained using a Bayesian estimation method proposed by Bock and Mislevy (1982). If we specify a prior distribution $f(\theta)$ for the proficiency level θ , then we can obtain the posterior distribution of θ using its likelihood function:

$$f(\theta \mid u_1, u_2, ..., u_n) = L(u_1, u_2, ..., u_n \mid \theta) f(\theta) ;$$

or

$$f(\theta \mid u_1, u_2, ..., u_n) = \prod_{i=1}^{n} (P_i^{u_i} Q_i^{1-u_i}) f(\theta) .$$

With this distribution, there are different statistics that can be use to estimate the value of θ . Bock and Mislevy proposed to use the mean and to call it the *expected a posteriori* estimate (EAP). They also proposed a method to estimate the *a priori* distribution $f(\theta)$. The θ estimates are obtain by finding the solution to:

$$E(\theta \mid u_1, ..., u_n) = \frac{\int_{-\infty}^{+\infty} \theta f(\theta) \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} d\theta}{\int_{-\infty}^{+\infty} f(\theta) \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} d\theta} .$$

The Hermite-Gauss quadrature is used to approximate the distribution $f(\theta)$ and the quadrature form of the preceding equation becomes (Baker, 1992, p.211):

$$E(\theta \mid u_1,\ldots,u_n) = \frac{\sum_{k=1}^{q} X_k L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)},$$

where X_k is one of q equidistant quadrature points comprised between θ = -4 and θ = 4, $A(X_k)$ is the weight associated to each of the quadrature points according to a normal probability distribution with mean zero and variance one, and where $L(X_k)$ is the likelihood of the responses pattern, r_j after the administration of j items. The following constraint is additionally imposed:

$$\sum_{k=1}^{q} A(X_k) = 1.$$

The integration was realized according to Mislevy's histogram method (Baker, 1992, p. 187), with 40 quadrature points and with weights equal to the a priori probability for these points (Bock and Mislevy, 1982; de Ayala, Shafer and Sava-Bolesta, 1995).

The standard error of the expected a posteriori estimate is calculated according to:

$$S_{EAP} = \left[\frac{\sum_{k=1}^{q} (X_k - EAP(\theta))^2 L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)} \right]^{1/2}.$$

The following equations, in concordance with the calculation of the centered moments proposed by Spiegel (1961, p.90), are respectively used to estimate the skewness a_3 and the kurtosis a_4 of the sampling distribution of the proficiency level estimate. The values of these statistics are used to determine if the shape of the sampling distribution of the estimate is symmetrical and not too far from a normal probability distribution. If these values are lower than 0.40, then we can be confident that the studied distribution is sufficiently close to a normal distribution to build a classical confidence interval with level α around the proficiency estimate (Raîche 2000, pp.146-158).

$$a_3 = \left[\frac{\sum_{k=1}^{q} (X_k - EAP(\theta))^3 L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)} \right] \Big/ S_{EAP}^3 \ ,$$

$$a_4 = -3 + \left[\frac{\sum_{k=1}^{q} (X_k - EAP(\theta))^4 L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)} \right] \Big/ S_{EAP}^4 \ .$$

In the simulation, all the tests end after the administration of 60 items. Nevertheless, the availability of the intermediary results (from 1 to 60 items administered) enable to know the value of the proficiency level estimate and its standard error after the administration of each of the 60 items. As a consequence, the results concerning that estimate after reaching a pre-determined level of the standard error of the proficiency level estimate are also available. The final proficiency level estimate is then equal to the provisory proficiency level estimate for the jth item administered.

3. Results

3.1 Maximum standard error as the stopping rule

Table 1 presents a summary of our observations as for the minimums and maximums of the statistics studied by relying upon the values of standard error ranging from 0.20 to 0.80. We can observe that when the standard error retained for the stopping rule is comprised between 0.40 and 0.20, the value of the skewness does not exceed 0.29 in absolute value, whereas that of the kurtosis does not reach beyond 0.39 in absolute value. We can also observe that the proficiency level estimate well covers the whole width of the integration interval and that the standard error of the proficiency level estimate has a difference of only 0.03 when the standard error retained for the stopping rule is les or equal to 0.40.

In general, the smaller the standard error retained for the stopping rule, the more the sampling distribution of the proficiency level estimate behaves according to a normal probability distribution. When the retained standard error for the stopping rule is equal or inferior to 0.40 it is not at all necessary to take into account the skewness and the kurtosis of the sampling distribution of the proficiency level estimate; these ones showing values within the interval ranging from –0.27 to 0.29, for the skewness, and from –0.39 to 0.23, for the kurtosis. These values affect only very little the standard interpretations related to a normal distribution N(EAP, S_{EAP}). For larger standard errors, the skewness and the kurtosis are located in a range of acceptable values but the EAP estimates do not span the entire integration interval and could produce some biased values for the extreme proficiency level.

According to certain supplementary analyses, we find it necessary to apply the correction suggested by Bock and Mislevy (1982) so as to alleviate the bias of the proficiency level estimate all over the proficiency level. It must be indicated, however, that that correction brings back the bias of the proficiency level estimate practically to zero only over part of the limited proficiency level to the interval [-3.00, 3.00] and in the case when the standard error retained for the stopping rule is below or equal to 0.40 (see Raîche and Blais, 2002).

Table 1: Minimums and maximums of the statistics associated with the sampling distribution of the proficiency level estimate with the stopping rule according to the standard error.

S.E.	**EAP estimate**		**S_{EAP}**		**a_3: skewness**		**a_4: kurtosis**	
	Min.	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**
0.85	-0.56	0.56	0.82	0.82	-0.11	0.11	0.13	0.13
0.80	-0.99	0.99	0.69	0.73	-0.18	0.18	0.16	0.22
0.75	-0.99	0.99	0.69	0.73	-0.18	0.18	0.16	0.22
0.70	-1.33	1.33	0.62	0.69	-0.22	0.22	0.17	0.22
0.65	-1.63	1.63	0.57	0.62	-0.24	0.24	0.15	0.24
0.60	-1.90	1.90	0.54	0.59	-0.24	0.24	0.12	0.25
0.55	-2.36	2.36	0.49	0.55	-0.25	0.25	-0.01	0.26
0.50	-2.76	2.76	0.45	0.50	-0.29	0.29	-0.20	0.26
0.45	-3.10	2.78	0.41	0.45	-0.26	0.28	-0.37	0.24
0.40	-3.38	3.05	0.37	0.40	-0.27	0.29	-0.39	0.16
0.35	-3.28	3.04	0.33	0.35	-0.22	0.23	-0.28	0.16
0.30	-3.28	3.21	0.29	0.30	-0.19	0.17	-0.15	0.17
0.25	-3.10	3.34	0.24	0.25	-0.15	0.15	-0.06	0.15
0.20	-3.16	3.16	0.20	0.20	-0.12	0.10	0.04	0.12

3.2 Number of items administered as the stopping rule

Table 2 presents a synthesis of our results for the stopping rule according to the number of items administered. The various values of the number of items administered retained for the stopping rule enable to approximate the most current situations observed in the literature about the adaptive tests and about the practice of their administration. The values of the skewness and of the kurtosis associated with the sampling distribution of the proficiency level estimate do not respectively exceed 0.29 and 0.41 in absolute value: values that affect very little the interpretations concerning the sampling distribution of the proficiency level estimate. With only ten items, however, the proficiency level estimate does not cover adequately the values of the integration interval.

We suggest that the stopping rule according to the number of items administered should be applied only if at least 13 items are administered. We also recommend the application of the correction proposed by Bock and Mislevy (1982) to bring the bias of the proficiency level estimate down to zero as far as the proficiency level is comprised within the interval [–3.00, 3.00] and that the number of items administered is at least equal to 10.

Table 2: Minimums and maximums for the statistics associated with the sampling distribution of the proficiency level estimate when the stopping rule according to the number of items administered is used.

Stopping rule	**EAP estimate**		**S_{EAP}**		**a_3: skewness**		**a_4: kurtosis**	
Item	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**
1	-0.56	0.56	0.82	0.82	-0.11	0.11	0.13	0.13
2	-0.99	0.99	0.69	0.73	-0.18	0.18	0.16	0.22
3	-1.33	1.33	0.61	0.67	-0.22	0.22	0.17	0.24
4	-1.63	1.63	0.55	0.62	-0.24	0.24	0.15	0.25
5	-1.90	1.90	0.50	0.59	-0.25	0.25	0.12	0.26
6	-2.14	2.14	0.46	0.56	-0.27	0.27	0.06	0.26
7	-2.36	2.36	0.43	0.53	-0.28	0.28	-0.01	0.26
8	-2.57	2.57	0.40	0.51	-0.29	0.29	-0.10	0.26
9	-2.76	2.76	0.38	0.49	-0.29	0.29	-0.20	0.25
10	-2.94	2.94	0.36	0.47	-0.29	0.29	-0.30	0.24
11	-3.10	2.78	0.35	0.46	-0.29	0.28	-0.37	0.24
12	-3.25	2.92	0.33	0.45	-0.29	0.28	-0.41	0.24
13	-3.38	3.05	0.32	0.42	-0.28	0.29	-0.39	0.24
14	-3.27	3.13	0.31	0.38	-0.25	0.23	-0.34	0.21
15	-3.38	3.25	0.30	0.38	-0.22	0.23	-0.36	0.39
20	-3.35	3.29	0.26	0.34	-0.21	0.23	-0.26	0.18
25	-3.16	3.35	0.23	0.28	-0.19	0.17	-0.15	0.14
30	-3.10	3.29	0.21	0.26	-0.16	0.16	0.00	0.11
40	-3.13	3.20	0.19	0.23	-0.12	0.11	0.04	0.08
60	-3.40	3.16	0.15	0.17	-0.10	0.08	0.01	0.06

Conclusion

The quest for precision in the measurement venture is not a new thing in the history of science as was illustrated in the book «The values of precision» edited by M.N. Wise (1995). This is precisely what constitutes one of the goals of computerized adaptive testing: being able to match more adequately the items and the examinees so that the estimate of the proficiencies are more informative. But there will always be some imprecision in measurement, since no imprecision would mean absolute precision, something impossible because it would also mean obtaining an infinite amount of information (Brillouin, 1964). According to Berka (1983, p.198), the imprecision of measurement depends on various factors: whether the probability distribution of the measure condenses around a certain value; whether it converges to a relatively stable value of deviation; whether it's dispersion is too wide to be confident about the inference on the center of the distribution.

These are precisely the features we have studied for a computerized adaptive test using the Rasch model and the expected a posteriori estimation method to estimate the proficiency level. With the Rasch model and the EAP estimation method, we found that if we want to postulate that the theoretical distribution of the proficiency estimate is normal N(EAP, S_{EAP}), then the stopping rule according to the number of items administered should be applied only if at least 13 items are administered and that, in general, it is preferable not to utilize the stopping rule according to the standard error with a standard error retained above 0,40. We also suggest the application of Bock and Mislevy’s correction (see Raîche and Blais, 2002), no matter which stopping rule is utilized, so as to reduce considerably the bias of the proficiency level estimate over the entire proficiency range.

References

Baker, F.B. (1992). *Item response theory: parameter estimation techniques*. New York : Marcel Dekker.

Berka, K. (1983). *Measurement: Its concepts, theories and problems*. Boston, Mass: Reidel.

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

Brillouin, L. (1964). *Scientific uncertainty and information*. New York: Academic Press.

Chang, H.H., Ying, Z. (1999). A stratified multistage computerized adaptive testing. *Applied psychological measurement*, 23, 211-222.

Chen, S.K., Hou, L., Dodd, B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and psychological measurement*, 58, 569-595.

Chen, S.K., Hou, L., Fitzpatrick, S.J., Dodd, B.G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and psychological measurement*, 57, 422-439.

de Ayala, R.J. (1992a). The influence of dimensionality on CAT ability estimation. *Educational and psychological measurement*, 52, 513-528.

de Ayala, R.J., Schafer, W.D., Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British journal of mathematical and statistical psychology*, 48, 385-405.

Dodd, B.G., de Ayala, R.J., Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.

Dodd, B.G., Koch, W.R., de Ayala, R.J. (1993). Computerized adaptive testing using the partial credit model effects of item pool characteristics and different stopping rules. *Educational and psychological measurement*, 53, 61-77.

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

Harwell, M.R., Stone, C.A., Hsu, T.C., Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, 20, 101-125.

Lord, F.M. (1980b). Some how and which for practical tailored testing. *In* L.J.T. van der Kamp, W.F. Langerak et D.N.M. de Gruijter (éds): *Psychometrics for educational debates*. New York: John Wiley and Sons.

Mislevy, R.J., Chang, H.H. (2000). Does adaptive testing violate local independence ? *Psychometrika*, 65, 149-156.

Nicewander, W.A., Thomasson, G.L. (1999). Some reliability estimates for computerized adaptive tests. *Applied psychological measurement*, 23, 239-247.

Raîche, G. (2000). La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur type et selon le nombre d'items administrés. Unpublished doctoral thesis. Montreal: University of Montreal.

Raîche, G., Blais, J-G. (2002). *Practical considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive correction for bias, and adaptive integration interval*. Paper presented at the 11th International Objective Measurement Workshop, Nouvelle-Orleans, April.

Spiegel, M.R. (1961). *Theory and problems of statistics*. New York : McGraw-Hill.

Thissen, D., Mislevy, R.J. (1990). Testing algorithms. In H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (éds.): *Computerized adaptive testing - A primer*. Hillsdale: Lawrence Erlbaum Associates.

Urry, V.W. (1970). *A Monte Carlo investigation of logistic mental models*. Unpublished doctoral thesis. West Lafayette: Purdue University.

van Krimpen-Stoop, E.M.L.A., Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied psychological measurement*, 23, 327-345.

Wise, M.N. (ed.)(1995). *The values of precision*. Princeton, NJ: Princeton University Press.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and psychological measurement*, 57, 412-421.