SECCIÓN MONOGRÁFICA: TESTS ADAPTATIVOS

Psicológica (2000), 21, 115-120 .

Overview of the computerized adaptive testing special section

Vicente Ponsoda[*]

Universidad Autónoma de Madrid

This paper provides an overview of the five papers included in the *Psicologica* special section on computerized adaptive testing. A short introduction to this topic is presented as well. The main results, the links between the five papers and the general research topic to which they are more related are also shown.

Key words: computerized adaptive testing, applications of item response theory.

During the last few years both books and a journal's special issue have been devoted to computerized adaptive testing (CAT): Drasgow and Olson-Buchanan (1999), Sands, Waters and McBride (1997) and the special issue of Applied Psychological Measurement, published in September 1999. The second edition of the book by Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg and Thissen (2000) has just been released, and that by van der Linden and Glas (in press) is coming soon. The interest in the topic was also evident in the last National Council on Measurement in Education (NCME) and American Educational Research Association (AERA) meetings: in the 1999 meeting, 25% of NCME contributions were related to CAT (Meijer and Nering, 1999).

The Spanish contribution to this area has not be important up to now, but some progress has been made and interest is increasing. Last two Spanish conferences on methodology for the social sciences, in 1997 and

[*] Facultad de Psicología, Universidad Autónoma, Canto Blanco, 28049, Madrid, Spain. E-mail: Vicente.ponsoda@uam.es Acknowledgements: I want to express my gratitude to all the authors of this special section for accepting taking part in it. Our research is being funded by DGES, grant PB96-0052.

1999, had symposia specifically on CAT. Another example of this interest is the early book written by Renom (1993). Last year, a new book edited by Olea, Ponsoda and Prieto (1999) appeared. This special section is an additional proof of this interest.

CAT's history is not too long. First developments appeared in the early seventies, but truly operative CATs started to be administered in the nineties. The idea behind a CAT is quite forward: to apply to each examinee only those items useful to know his/her proficiency level. As a consequence of this, CAT is more efficient than conventional (i.e., fixed-item) tests. It provides more precise measurements for same-length tests or shorter tests for same-precision measurements. The basic elements of a CAT are an item pool, a procedure for ability estimation, a heuristics to select items and a stopping rule. Early CATs selected items based only on the information principle: the unused item most informative at the last ability estimate was selected and administered. The two most common stopping rules are a prefixed test length or standard error. The theoretical bases are provided by Item Response Theory (IRT). Its invariance property makes it possible to obtain ability estimates in the same scale despite examinees having received different set of items. IRT developments on item calibration, ability estimation, item pool dimensionality,.. are in use in CAT. Wainer et al. (2000), Renom and Doval (1999) or Olea and Ponsoda (1996), these last two in Spanish, may be good introductory sources to the topic.

Current CAT research is dealing with topics such as these: a) extending CAT to non-dichotomous items (politomous and constructed-response items), b) multidimensional CAT, c) adding restrictions to optimal selection criteria (item exposure control, alternative information measures to guide item selection, content and other constraints,…), d) examinee issues (item review,…), e) adaptive and sequential mastery testing, and f) others (item differential functioning, aberrant response patterns, multi-stage and modularized adaptive testing, response times, ability estimation in CAT, integrating assessment into learning and diagnosis systems,…). The special issue is comprised of five papers. The first two will be mostly concerned with the interesting practical problems raised by current operational CATs. The remaining three papers will consider issues referred to above in the previous list. Each one will be commented on in the following paragraphs.

Wainer's contribution (CATs: Whither and whence) compares the prospects Bert Green gave to CAT in1983 with its current status. It seems to me that Dr. Wainer has been in doubt about whether giving a positive or negative vision for the future of CAT. However, his final vision is to some extent positive. As the paper shows, during the last decade most than three million CATs have been administered. One main question of the paper is

what we should learn from this impressive experience. We have learnt the importance of item exposure control, the need to implement time limits for examination (despite that in theory the examinee should not have such a time limit), we have also learnt the difficulties of keeping a test secure when it is continuously applied (more on this in the contribution by Wise and Kingsbury in this issue), and the huge costs involved in developing and maintaining operational CATs. The paper also reviews the advantages paper and pencil testing continues having today: low cost, more places for testing (because there is no need for special testing centres), and no problems for answer revision (more on this last topic in Olea, Revuelta, Ximénez and Abad, and Wise and Kingsbury, in this issue). The paper's main conclusion is that CAT is good for some applications, but not so good for others. CATs should be considered when the construct to be measured needs or may benefit from the use of computers (i.e., architectural design tests), when the test has to be offered continuously in time (i.e., when an examination delay implies an extra cost for the examinee), but CATs are not the best option for exams that may be applied one or two times per year, as entrance exams. In these cases continuous testing is an undesired consequence of CAT (due to the shortage of CAT sites), and not a desirable feature of the exam. However, some big programs using CAT are entrance exams. As said above, Dr. Wainer's final idea is that CAT is promising for specific tests, not for all.

Prof. Wise and Dr. Kingsbury's contribution (Practical issues in developing and maintaining a computerized adaptive testing program) gives their view on the difficulties arising in planning, implementing and maintaining a CAT program. The paper has four main areas: item pools, test administration, test security, and examinee issues. Concerning items pools, the paper stresses the more and more obvious need for greater item pool sizes, as a result of the increased concerns on test security and the more and more sophisticated item selection rules in use, that incorporate content balancing and item control exposure restrictions. Recommendations are given on the IRT model to choose, uni or multidimensional, on how to exclude or add items to the current item pool, and on the use of multiple item pools. Specially revealing is the discussion about how to keep scale consistency under control. The second part deals with test administration. Test entry and termination strategies, item selection and scoring procedures are revised. The third part, on test security, focuses on the impact of item disclosure and item theft on the psychometric properties of CATs. The last part of the paper deals with examinee issues, in which Prof. Wise has strong interest. CAT differs in some clear ways from paper and pencil testing and the idea is to consider the psychological consequences these differences may have for the examinees. Three particular issues are considered: Item review,

time limits and equity. Item review is the specific topic of Olea et al. in this issue. The determination of time limits is a hard task in CAT and the possible errors and their consequences are outlined. Some considerations on factors that may compromise score comparison between examinees are also provided. The paper dedicates its final comments to the areas in which research should focus in the near future.

Olea, Revuelta, Ximénez and Abad (Psychometric and psychological effects of review on computerized fixed and adaptive tests) show us a well-conducted piece of research on the effects of review on psychological and psychometric variables, both in fixed and adaptive tests. The paper comments on the pros and cons of review in CAT. Test agencies and administrators do not seem yet convinced by these pros, as the paper acknowledges that only one operative CAT has implemented some sort of item review up to now. The results from the experiment show that state-anxiety level decreased in the review condition whereas it slightly increased in the non review condition. A high percentage of people allowed to review actually did so; and, in concordance with previous studies, most of them received a higher ability estimate after revision. The legitimacy of score gains due to item review is considered in some detail. Item review is a controversial issue. Wainer's paper in this issue summarises well the position against the use of review: CAT's main achievement is efficiency and this is reduced if revision is implemented. On the other hand, examinees clearly prefer to review their answers, as they can do in conventional paper and pencil testing, and doing so they quite often improve their scores. The paper ends with a positive advice concerning review implementation in CAT.

Using computers for test administration makes it possible to gather item response times simultaneously with response correctness. Prof. Hornke's paper (Item response times in computerized adaptive testing) considers what additional meaning we may get from those response times. One important problem is whether or not response times give information on the same ability we want to measure or rather on other construct (such as a personality trait). The paper shows that both old references (some of them from the early 30s) and newer cognitive modelling approaches acknowledge both possibilities. The paper gives data on an empirical study based on a unusual big sample of more than 5000 people taking an adaptive matrices test. Response time means (and variances) are higher for wrong responses than for correct ones. The increase in time needed for wrong responses has been reported a few times before. Pearson correlations between ability estimates and response times are in the 0.50 - 0.65 range. However, there are hints in the results indicating that the relationship should be better traced by a non-linear correlation index. Psychological processes involved in

correct and incorrect responses may be different. If this is so, response times may also indicate those distinct processes and should show different correlation patterns with other tests. This was not the case in the study. Correlations between response times and a wide range of aptitude tests were the same for correct and for incorrect responses. No conclusive answer could be reached on the exact meaning response times have.

The paper by Dr. Vos (A bayesian procedure in the context of sequential mastery testing) closes the special section. In mastery testing the final decision for the examinee is a final category (i.e., pass or fail) rather than an ability estimate. This testing is specially suited for certification, licensure or graduation purposes. The paper distinguishes between different types of mastery testing. There are fixed- and variable-length mastery tests. Variable-length mastery tests may be adaptive or sequential. In the first case, item selection depends on the examinee's ability; on the contrary, in the second case, items are selected at random. Sequential mastery testing can be studied in the framework of Bayesian sequential decision theory. Two are its basic elements: the psychometric model, relating the unknown examinee's ability level to his/her probability of correct response, and the loss function. Under this approach, optimal rules may be obtained. In the paper, the binomial function is taken as the psychometric model, a simple threshold loss function is used, and optimal rules are obtained applying dynamic programming techniques. The binomial function gives the probability of "k" correct responses when "n" items have been administered to an examinee with a certain ability level. The loss function specifies at each testing stage the costs and gains of each possible decision for examinees above and below a cut score. Optimal rules specify the option for which posterior expected losses are lowest. The efficiency of the proposed procedure is compared by simulation to fixed-length and other types of sequential mastery testing. Results are encouraging. Some comments on recent extensions of the approach are also offered.

RESUMEN

Presentación de la sección monográfica sobre test adaptativos informatizados. Este artículo proporciona una visión conjunta de la sección especial de *Psicológica* sobre tests adaptativos informatizados. Se presenta también una breve introducción al tema. De cada artículo se muestran sus principales resultados, las conexiones con los demás trabajos de la sección especial y el tema de investigación con el que está más relacionado.

Palabras clave: test adaptativos informatizados, aplicaciones de la teoría de respuesta al ítem.

REFERENCES

Drasgow, F. and Olson-Buchanan, J.B. (Eds.) (1999). *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

Meijer, R.R. and Nering, M.L. (1999). Computerized adaptive testing: Overview and Introduction. *Applied Psychological Measurement*, 23 (3), 187-194.

Olea, J., Ponsoda, V., and Prieto, G. (Eds.) (1999). *Tests informatizados. Fundamentos y Aplicaciones*. (Computerized tests. Foundations and applications). Madrid: Pirámide.

Renom, J. (1993). *Tests adaptativos computerizados. Fundamentos y aplicaciones*. (Computerized adaptive testing: Foundations and applications). Barcelona: PPU.

Renom, J. and Doval, E. (1999). Tests adaptativos informatizados: Estructura y desarrollo. (Computerized adaptive testing: Structure and implementation). In Olea, Ponsoda and Prieto (Eds.) (1999). *Tests informatizados. Fundamentos y Aplicaciones*. (Computerized tests. Foundations and applications). Madrid: Pirámide.

Sands, W.A., Waters, B.K. and McBride, J.R. (Eds.) (1997). *Computerized adaptive testing. From Inquiry to operation*. Washington: American Psychological Association.

van der Linden, W.C. and Glas C.A.W. (Eds.) (in press). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer-Nijhoff.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L. and Thissen, D. (2000). *Computerized adaptive testing: A primer*. 2nd edition. Hillsdale, N.J.: Erlbaum.