Computerized adaptive testing for measuring development of young children

Gert Jacobusse[1,*,†] and Stef van Buuren[1,2]

[1]*TNO Quality of Life, Leiden, The Netherlands*
[2]*Utrecht University, Utrecht, The Netherlands*

SUMMARY

Developmental indicators that are used for routine measurement in The Netherlands are usually chosen to optimally identify delayed children. Measurements on the majority of children without problems are therefore quite imprecise. This study explores the use of computerized adaptive testing (CAT) to monitor the development of young children. CAT is expected to improve the measurement precision of the instrument. We do two simulation studies—one with real data and one with simulated data—to evaluate the usefulness of CAT. It is shown that CAT selects developmental indicators that maximally match the individual child, so that all children can be measured to the same precision. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: computerized adaptive testing (CAT); child development; developmental monitoring; Rasch model; measurement

1. INTRODUCTION

Children develop a wide range of skills that are crucial for their lifelong health and quality of life. Skills that are not acquired at the usual age may result in a delayed development that is hard to catch up. It is therefore important to monitor development very closely, and to identify any delay before it has irreversible consequences. Identifying developmental problems in very young children needs special attention; a US study showed that children younger than three years are least likely to use health services. This finding held when their reduced disease risk was taken into account [1].

---
*Correspondence to: Gert Jacobusse, Department of Statistics, TNO Quality of Life, P.O. Box 2215, 2301 CE Leiden, The Netherlands.
†E-mail: gert.jacobusse@tno.nl

In The Netherlands, the Van Wiechen Scheme [2, 3] is routinely applied to monitor the development of children from birth up to four years of age. This scheme consists of developmental indicators that inquire the presence of various behaviours, for example 'reacts to speech' or 'sits without support'. The scheme is administered by a health professional, typically a physician or a trained nurse. For each indicator, a 'pass' score signals that the particular behaviour is present, a 'fail' score otherwise. Indicators can be ordered according to difficulty, e.g. standing is more difficult than sitting. Indicators are divided into sets that increase in difficulty and that are meant for children of a certain age.

Age specific sets in the original Van Wiechen scheme were constructed such that approximately 90 per cent of the children of the target age achieve a pass score on each of the indicators. So a child failing on one or more indicators is relatively unexpected. The information that is gathered in this way is very useful to identify delayed development. On the other hand, the information about normal development of the majority of children, who pass all indicators, is minimal. The only possible conclusion is 'no problem identified'.

It is sometimes beneficial to have a more refined measurement of development, for example, when developmental progress in time is studied. A better measure is possible if a larger number of indicators of different difficulty is administered. Having both fail and pass scores results in a more precise estimate of the developmental score or 'D-score' that can be derived from the responses to a series of indicators. The disadvantage, of course, is that the evaluation of a larger set of indicators costs more time.

An alternative strategy for more refined measurement is to have a small customized set of indicators that is chosen optimally for each individual child. Computerized adaptive testing (CAT) is a technique to find this optimal set of indicators [4]. In CAT, the computer calculates a preliminary estimate of the D-score of the individual child after every result that is entered. The next indicator is chosen so that its difficulty maximally matches the D-score of the child. Although each child is measured by means of a specific subset of indicators, the D-score estimates that the model reveals are on the same scale and can be used to compare children across tests, individuals, or time [5]. We evaluate the usefulness of CAT to select Van Wiechen Scheme indicators and measure the development of young children. CAT is expected to improve the measurement precision of the instrument.

2. METHODS

2.1. Underlying model

The Rasch model [6] describes the probability to achieve a pass score as a function of the D-score of a child and the difficulty of an indicator. Both the D-score and the difficulty can be expressed as positions on a latent scale that represents the level of development. The D-score of child i $(i = 1, \ldots, n)$ has position θ_i and the difficulty of indicator j $(j = 1, \ldots, m)$ has position δ_j. The Rasch model describes the probability that child i passes indicator j as

$$P(X_{ij} = 1 | \theta_i, \delta_j) = \exp(\theta_i - \delta_j)/\{1 + \exp(\theta_i - \delta_j)\} \qquad (1)$$

The probability to fail $P(X_{ij} = 0)$ is equal to $1 - P(X_{ij} = 1)$. For the application of CAT, the item difficulties δ_j are known, and the pass/fail scores X_{ij} are available for indicators that already have

been administered. In order to estimate the D-score positions θ_i, we can use Bayes rule [7] as

$$P(\theta_i|\delta_j, X_{ij}) = P(\theta_i) * P(X_{ij} = k|\theta_i, \delta_j) \qquad (2)$$

where $P(\theta_i)$ is the prior distribution of θ_i, and $P(\theta_i|\delta_j, X_{ij})$ is the posterior distribution of θ_i after child i obtained score k on item j. After rescaling the posterior distribution $P(\theta_i)$ to unit area it is used as the new prior distribution for the next indicator. The centre of gravity of the posterior distribution is called the expected *a posteriori* (EAP) estimate, or expected value of $P(\theta_i|\delta_j, X_{ij})$. If the X_{ij} follow the model in (1), this estimate is expected to get more precise as $P(\theta_i|\delta_j, X_{ij})$ narrows after every new result.

2.2. CAT procedure

During a CAT session, selection of the next indicator for a particular child is based on the difficulties δ_j of the indicators still left in the indicator pool, and the child's (updated) prior distribution $P(\theta_i)$. Before any indicator is answered, an initial prior distribution is needed for selecting the first indicator and calculating a posterior distribution. The current practice of the Van Wiechen scheme is to select those age-specific indicators that roughly match the child's level of development. Analogous to that, we use the age-conditional distributions of D-score position θ_i and specify the child's prior as a normal distribution, with the mean equal to the expected value given the child's age, and with a standard deviation of 4—which is over dispersed, about two times the expected standard deviation.

Within the CAT algorithm, the indicator is selected that maximally matches the current D-score of the child. A good match implies that the response to the indicator gives a large amount of new information. The less predictable the response, the more information it contains. For a binary indicator, a response is least predictable if the probability of passing is equal to the probability of failing. This is the case when D-score position θ_i and indicator difficulty δ_j are equal. In that case, $P(X_{ij} = 0)$ and $P(X_{ij} = 1)$ are both 0.5. The indicator whose difficulty δ_j is closest to the EAP estimate of θ_i is therefore selected as the next one. Note that this is identical to selection by the Fisher information measure [8].

In this way, the CAT algorithm selects one indicator at a time from all available indicators, administers it, and removes the selected indicator from the pool of available indicators. The algorithm can stop in one of the following circumstances:

   after a fixed number of indicators has been administered;
   after the estimation of θ_i is deemed sufficiently precise;
   after the pool is fully exhausted.

The precision of θ_i depends on the measurement error, which is the width of the posterior distribution. In this study, we stop the algorithm after a fixed number of indicators. This allows us to compare the precision of the CAT method to that of the conventional age-specific set having the same number of indicators.

2.3. Simulation study

To evaluate the usefulness of CAT for selection of Van Wiechen Scheme indicators, we run a simulation study on data that were collected within the 'Social medical survey of children attending child health clinics' [9]. This is a longitudinal study on 2151 children. The data set is
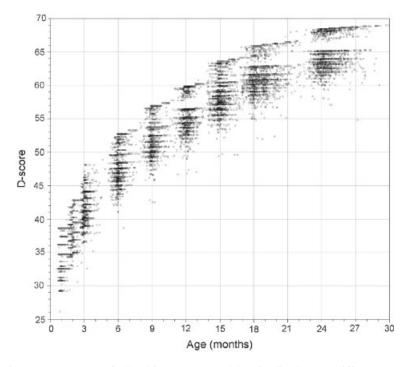
Figure 1. *D*-scores of all subjects, measured longitudinally at 9 different ages.

unique, because age-specific indicators for the target age of the child were administered together with indicators that are targeted at slightly older children.

Estimation of the indicator difficulties and *D*-scores according to the Rasch model was carried out using the RUMM [10] software, and is described by Jacobusse *et al.* [11]. The Rasch model fits the data quite well, the estimated reliability (person separation index) for the whole scale is 0.99 and outfit mean square statistics of the 12 indicators used in this study vary between 0.53 and 1.17. Most important, *D*-scores have almost perfectly parallel associations with percentage pass scores for all indicators.

The *D*-scores of all subjects at different ages are given in Figure 1. The figure displays Bayesian EAP estimates that are usual in a CAT context. They slightly differ from the weighted likelihood estimates in the original article. For the simulation study, we select only the cases at about 12 months, the fourth data cloud from the right. A set of 12 indicators was evaluated at this age. Six of them were targeted at children of 12 months, the other six at children of 15 months. Locations of the 12 indicators are given in Table I.

In the first simulation study, three different kinds of *D*-score estimates will be compared to the original *D*-scores in Figure 1 that use the full 12 indicators. The simulation study will only use the 1221 children whose results on all 12 indicators are known. Three different indicator sets will be considered. First, we derive an estimate based on the 6 indicators targeted at children of 12 months (Method A). Second, we estimate the *D*-score based on the 6 indicators targeted at children of 15 months (Method B). Third, we use the CAT procedure described in Section 2.2 and stop after 6 indicators, either targeted at month 12 or month 15, have been administered (Method C).

Table I. Locations of the 12 indicators.

Item formulation	Location (D-score scale)
Target age 12 months	
Sits without support	49.9
Picks up crumb between thumb and index finger	51.7
Crawls	51.7
Pulls himself to standing position	52.4
Waves 'bye bye'	51.7
Jabbering	50.4
Target age 15 months	
Gets cube into and out of box	53.3
Plays 'give and take'	53.6
Crawls, with belly lifted off the ground	53.4
Walks while holding furniture	53.4
Understands some simple words	53.2
Uses two words	55.6

A second simulation study will be carried out on simulated data. First, we randomly generate 1000 D-scores from a uniform distribution. After that, responses to all indicators—including those targeted at children younger than 12 and older than 15 months—are generated according to the Rasch model in (1). The CAT procedure is then used to select either 3, 6 or 12 items from the complete indicator pool. Results of this simulation study will show how availability of all indicators and variation in the number of indicators take effect.

3. RESULTS

Within the fixed subsets of indicators, i.e. in methods A and B, the D-score follows directly from the total score; this is the number of pass scores. Table II provides a key to translate total scores into D-scores, using the prior for children aged 12 months.

Method C dynamically selects 6 out of 12 indicators. This may result in $12!/(6! * (12-6)!) = 924$ different subsets, so a table for translating all possible total scores into D-scores would take a few pages. Fortunately, there is no need for such a table—the CAT algorithm applies (2), and directly returns the D-score as a result.

Figure 2 plots the D-score estimates under methods A, B and C against the 'true' D-scores based on all 12 indicators. The D-score estimates from method A match the true scores only for delayed children, i.e. at the left of the figure. This is a direct consequence of the choice for indicators with a 90 per cent passing rate. D-score estimates by method B are quite good for children with a true score of around 54, but more variable in the extremes, leading to the diabolo-like shape.

D-scores by method C are as good as method A in the low scores, as good as method B for the middle scores, and still not so good for children with higher true scores. This indicates that the CAT algorithm selected the right questions to optimally measure each individual child.

In the second simulation study, the CAT algorithm was applied to select 3, 6 or 12 indicators from the complete indicator pool. Figure 3 shows that children with different true D-scores can all be measured with the same precision. If enough indicators of diverse difficulty are available, the precision of the estimate only depends on the number of indicators that is administered.

Table II. Key to translate total scores into *D*-scores for children aged 12 months.

Total score subset 12 months	*D*-score	Total score subset 15 months	*D*-score
0	46.8	0	48.7
1	49.3	1	51.5
2	50.5	2	52.8
3	51.5	3	53.8
4	52.5	4	54.8
5	54.0	5	56.2
6	58.2	6	59.7

4. DISCUSSION

We have shown that CAT for measuring the development of young children considerably improves estimates of *D*-score. Unlike the traditional fixed age-specific sets of indicators, CAT dynamically selects a set of indicators that maximally matches the development of the individual child. Practical application of CAT for measuring development in young children increases precision of *D*-score estimates, and thereby improves upon monitoring of development.

The CAT technique originates from educational research, and is becoming more and more accepted in different fields of health care, like psychological measurement [12] and pediatric rehabilitation [13]. The present research illustrates the potential of CAT for measurement of child development, where application of CAT is relatively new.

4.1. Methodological considerations

The first simulation study uses a real data estimate that is based on 12 indicators as the 'true' *D*-score. This true score has its limitations. Estimated scores (Figure 2) are based on a subset of the same indicators that also underlie the true scores. Thus, the information that is contained in the pass/fail scores X_{ij} is used twice, once for the true scores and once for the estimates. This explains why the range of estimated scores given true scores is sometimes smaller in Figure 2 than what would be expected from the 6-item CAT in Figure 3. The highest true scores in Figure 2 represent children who passed all 12 indicators—and therefore all indicators in any subset of 6. Especially for these children, both the variation in true scores and the variation in estimated scores are limited by the indicators that were included in the data collection design.

The second simulation study addresses the limitations of the first. The true *D*-scores are simulated, and the pass/fail scores X_{ij} depend on the true scores only through the Rasch model in (1). Further, results on all indicators are known, so that the variation in estimated scores is no longer limited by the availability of indicators. The disadvantage of simulated data is, however, that the assumed model always fits the data perfectly. Together, the two simulation studies give a more complete picture of the usefulness of CAT.

The Rasch model assumes that the underlying scale is unidimensional. Indicators in the Van Wiechen Scheme represent different areas of development like motor, psychological and language skills. This variation in traits may be a problem in the context of CAT if indicators from the same area cluster around a certain *D*-score level. However, this does not seem to be the case here. Indicators in the original Van Wiechen Scheme are divided into age-specific sets of approximately
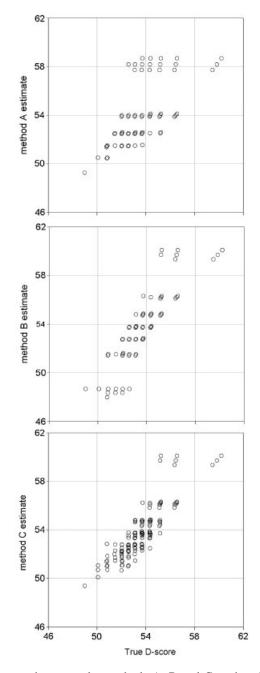
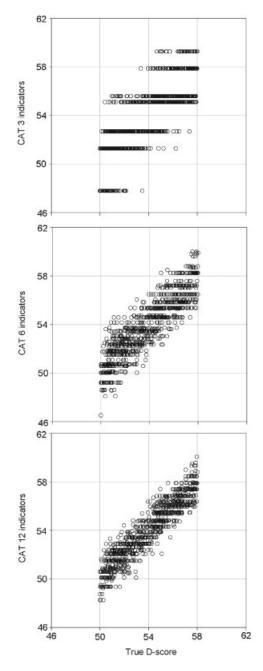Figure 2. *D*-score estimates under methods A, B and C against 'true' *D*-scores.

Figure 3. CAT estimates after 3, 6 and 12 indicators against true *D*-scores.

the same difficulty, and each set contains indicators that represent the whole developmental spectrum. It is therefore improbable that a CAT will just select items that are restricted to one area of development. If more indicators are going to be added in the future, a kind of 'content balancing' may be useful to avoid that some children will be measured on just one area of development.

In a CAT like this with exclusively binary indicators following the 1 parameter Rasch model, exposure of items at a given D-score level only depends on the item location (indicator difficulty), because all information curves are of the same form. Items with locations that are close together have a smaller chance of being selected, because the range in which they have maximum information is smaller (i.e. there are more alternatives). But, in contrast to a CAT with mixed numbers of response categories or different discrimination parameters, there are no items that are never selected.

CAT is usually applied for self-report, i.e. questionnaires in which the respondent answers the questions. The Van Wiechen Scheme is administered by health professionals, and is thus somewhat unique as subject of CAT. The person-specific variation that is inherent to self-report is avoided in this setting, but this is traded in for a possible variation between professionals. We do not think that this is a major problem. Professionals are well-trained and see a lot of children. The latter provides an opportunity for further research, as differences in how professionals use the indicators could be formally tested.

4.2. Practical considerations

Before CAT can be routinely applied to monitor the development of children, some conditions need to be met. In the first place, the right software to perform CAT needs to be developed. This software should be easy to use, have an attractive interface, and immediately report the D-score once the session is stopped. Personal computers should be available on the work floor. The user of the software may need some training in the appropriate use of the software and in the interpretation of the outcomes. If such operational barriers are overcome, the technique presented in this paper allows for an efficient and precise way to quantify development.

REFERENCES

1. Stahmer AC, Leslie LK, Hurlburt M, Barth RP, Webb MB, Landsverk J, Zhang J. Developmental and behavioral needs and service use for young children in child welfare. *Pediatrics* 2005; **116**(4):891–900.
2. Schlesinger-Was EA. *Ontwikkelingsonderzoek van zuigelingen en kleuters op het consultatiebureau*. Leiden University: Leiden, 1981.
3. Den Ouden L, Rijken M, Brand R *et al*. Is it correct to correct? Developmental milestones in 555 'normal' preterm infants compared with term infants. *Journal of Pediatrics* 1991; **118**(3):399–404.
4. Wainer H. *Computerized Adaptive Testing*, *A Primer*. Lawrence Erlbaum Associates: New Jersey, 1990.
5. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for 21st century. *Annals of Internal Medicine* 1997; **127**:743–750.
6. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedogogische Institut: Copenhagen, 1960.
7. Van der Linden WJ, Glas CAW. *Computerized Adaptive Testing*, *Theory and Practice*. Kluwer Academic Publishers: Dordrecht, 2003.
8. Lord F. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum: Hillsdale, NJ, 1980.
9. Herngreen WP, Reerink JD, Noord-Zaadstra BM *et al*. SMOCC: design of a representative cohort-study of live-born infants in The Netherlands. *European Journal of Public Health* 1992; **2**:117–122.
10. RUMM Laboratories, 2005. www.rummlab.com.au.

11. Jacobusse GW, Van Buuren S, Verkerk PH. An interval scale for development of children aged 0–2 years. *Statistics in Medicine* 2006; **25**(13):2272–2283.
12. Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. *Medical Care* 2002; **40**(9):812–823.
13. Haley SM, Raczek AE, Coster WJ, Dumas HM, Fragala-Pinkham MA. Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation* 2005; **86**(5):932–939.