8 Item Response Theory based Computerized Adaptive Testing can provide an accurate and efficient identification of children with psychosocial problems

A.G.C. Vogels
G.W. Jacobusse
S.A. Reijneveld

Submitted
Abstract

Objectives
To assess if a Computerized Adaptive Test can overcome the weaknesses of short written questionnaires when identifying children with psychosocial problems.

Research design
We used an existing national data set with 205 items on behavioral and emotional problems (n=2041). In a first random subsample we determined which items met the requirements of an IRT model sufficiently. Using those items, item parameters necessary for a Computerized Adaptive Test were calculated and a cut-off point was defined. In the second subsample we determined the validity and efficiency of a Computerized Adaptive Test in a simulation study, with current treatment status and the Total Problem Scale of the Child Behavior Checklist as criteria.

Subjects
Respondents were 2041 parents of children invited for a routine health examination by Preventive Child Healthcare (PCH) (response: 84%).

Results
Out of 205 items available only 15 did not sufficiently meet the criteria of the underlying IRT model. For 90% of the children a score above or below cut-off point could be determined with 95% accuracy. Sensitivity and specificity with the Total Problem Scale as a criterion were 0.89 and 0.91. The mean number of items needed to achieve this was 12.

Conclusion
An IRT-based CAT is a very promising option for the identification of psychosocial problems in children, as it can lead to an efficient, yet high-quality identification. The results of our simulation study need to be replicated in a real-life administration of our Computerized Adaptive Test.
8.1 Introduction

Many children suffer from behavioral and emotional problems,\(^1\) and these problems may seriously interfere with their daily functioning, now and later in life.\(^4,5\) Yet many of these children remain untreated.\(^6\) Durlak and Wells showed that early identification considerably improves the prognosis of the children involved.\(^7\)

Community-based preventive child healthcare (PCH) services, especially those working outreachingly, are in a unique position to identify such problems as early as possible. In the Netherlands, PCH professionals offer routine well-child care to the entire Dutch population to the age of about 14. The early detection of children with psychosocial problems is an explicit part of their working package. In contrast to systems existing e.g. in the US, Dutch PCH does not offer treatment services. When (physical or psychosocial) problems are detected, children are referred to other parts of the healthcare system, especially to primary healthcare. Research has shown, however, that early identification in PCH is often far from perfect. For example, Brugman et al. showed that in Dutch PCH, about half of the children with a clinical CBCL Total Problem Score remained unnoticed when they were examined by a physician or nurse.\(^1\) Other studies came to similar conclusions.\(^8\)\(^-\)\(^11\)

There are several possibilities to improve the identification of children with emotional and behavioral problems. Wiefferink et al. showed that using clear protocols and extensive staff training can lead to a significant increase in the number of children with problems identified and a decrease in the number of children incorrectly identified as having problems.\(^12\) Other studies showed that using good questionnaires, to be filled in by parents, teachers or the children themselves, can also help to improve the quality of early identification.\(^7,13,14\) However, in community-based PCH, the time available for each individual child is limited. This means that questionnaires that are practicable in such settings, have to be easy to score and therefore short. Also, they must be easy for all parents to answer. Short questionnaires, unless they have a very narrow scope, tend to be less reliable and less valid than desirable.\(^15\) Identification of problems based on such questionnaires is therefore error prone, resulting in too many false classifications.

Since the 1950s, a series of new statistical models called Rasch or IRT (Item Response Theory) models have been developed which allow for Computerized Adaptive Testing (CAT), a short and efficient test procedure that does not compromise the accuracy of the test results. Originally these models could only be applied to items with only two categories. This limited their application mainly to the field of intelligence testing and the assessment of school achievements.\(^16\) In the last decades more widely applicable models became available. This led to IRT-based test procedures in the field of quality of life measurements.\(^17\) Very recently some publications have been published describing the application of these models to the assessment of mental health problems.\(^18\)\(^-\)\(^20\)

In this study we assessed whether CAT can also be used for a fast, short, yet high-quality identification of children with emotional and behavioral problems in community-based PCH.
8.2 Methods

8.2.1 General introduction
Just like test procedures based on more traditional psychometric theories, IRT-based procedures help to determine the position of a person on some measurement scale, for instance on intelligence, school achievements or the level of psychosocial problems. In IRT that position is called the person location. IRT differs from traditional psychometrics in that it provides information about which items are relevant to use in an individual assessment and which are less useful.

A simple example may illustrate this principle: suppose in a particular arithmetic test, a child failed to give the correct answer to the question: how much is 2*3. In that case it is probably not very useful to ask: how much is 34*17. The latter question can help to distinguish between children on a higher position of the arithmetic ability scale, but will add little information for a child who failed the first question. Translating this to scales assessing emotional and behavioral problems, items indicating severe problems are not informative for children with no or few problems and items indicating less severe problems are not informative for children with severe problems.

With IRT it is possible to determine the severity of individual items; i.e. the position on the scale where it is informative. That position is called the item location. This information can be used to shorten the test length in the following way. After each answer on a single question an estimation is made of the person’s probable score, or person location. Then the available items are scanned in order to determine which item could improve the estimated person location. This continues until a previously defined accuracy has been reached. In practice this process is only possible with the aid of computers: Computerized Adaptive Testing (CAT). For CAT to be possible, the location of the items must be known in advance, before actual testing of an individual starts.

In order to assess whether such a procedure is viable and efficient for the identification of children with problems, three questions need to be answered:

1. Are the items of four questionnaires on emotional and behavioral problems suitable for an IRT-based CAT and, if so, what are parameters of the individual items, to be used in a CAT?
2. Which cut-off point results in a sensitive and specific distinction?
3. What are the validity and the specificity of such a CAT and how efficient is this procedure?

8.2.2 Data collection, population and measures
We used a data set collected in an earlier study containing information about parent-reported problems of children aged seven to twelve. Data were collected in a two-step procedure. In the first step nine regional PCH organizations were found willing to participate in our study. Second, parents who were invited for a routine health examination of their child were asked to participate in the study and to fill in some questionnaires about emotional and behavioral disorders of their child.
Data from 2041 parents were available, 84% of the parents invited to participate. 51% of the children involved were girls. Mean age was 10.1 (sd=1.4). In 83% of all cases both parents were born in the Netherlands and in 6% at least one parent came from a non-OECD country or from Turkey. Fourteen percent did not live in a two-parent family and only three percent lived in a family were none of the parents was employed. The sample was representative for the population of this age group under care in Dutch PCH, regarding gender, age, ethnic origin and family composition (two parents, one parent, other), with Cohen’s W varying from .002 for gender to .109 for ethnic origin.

Each parent answered the Child Behavior Checklist (CBCL) and one out of three questionnaires: the Pediatric Symptom Checklist (PSC), the Strengths and Difficulties Questionnaire (SDQ), or a newly developed Dutch questionnaire on psychosocial problems for children in primary education, the PSYBOBA. More information on the relatively unknown PSYBOBA may be obtained from the authors. The PSC, SDQ and PSYBOBA were chosen for this study because there was evidence for their conceptual validity in relation to the kind of problems Dutch PCH aims to identify and because they met the requirements for use in the context of PCH: short, easy to administer and to score. The data collection led to an incomplete data matrix: the data for the PSC, the SDQ and the PSYBOBA are each available in about one third of the sample.

Finally, data on current treatment status for emotional or behavioral problems were obtained from the PCH professionals. These answered questions regarding this, based on medical records and on the routine health examination of the child, during which a small structured interview was done for the purpose of this study. One of the questions asked whether the child was currently being treated for mental health problems and, if so, by which kind of professional or institution.

8.2.3 Analyses
We randomly divided the total sample in two subsamples. The first one, the calibration sample (n=1650), was used to answer the first two questions (suitability of the items and determination of the cut-off point). The second, the validation sample (n=391), was used for the evaluation of the validity and efficiency. This evaluation in a separate sample was done in order to prevent overestimation of validity and efficiency coefficients.

In order to assess the suitability of the items for an IRT-based CAT we assessed whether the items fitted the assumption of unidimensionality. For this aim, we determined whether the items showed enough fit with the Partial Credit Model (PCM), one of the unidimensional IRT models. Using this model for a CAT has the advantage that it results in scores on an interval measurement level. We performed this assessment using the RUMM 2020 software (http://www.rummlab.com.au), as this can handle incomplete data matrices like ours. RUMM 2020 provides so called outfit statistics for each item, that indicate to what extent each item fits the model. Items were considered suitable for CAT measurement if they had an outfit statistic smaller than 1.7.

Next, we calculated the item locations of the remaining items, using the same software. In order to determine whether the estimated item locations would be valid, independently from gender and ethnicity, we performed Differential Item Functioning
(DIF) analyses for each item. We did this by multinomial logistic regressions, with the raw score on the item as the dependent variable. First, the estimated person location was the only predictor in the logistic regression model. Second, both gender, ethnicity and their interaction were added as predictors. Items were considered as showing DIF when these additional predictors had a significant effect and led to an increase of the explained variance of more than 3.5%.

Second, we determined an optimal cut-off point for the CAT scores, i.e. one which enables a good distinction between a non-clinical versus a clinical CBCL TPS. The CBCL TPS was used as the criterion measure, because it measures exactly the emotional and behavioral problems which Dutch PCH aims to identify and because both its concurrent and predictive validity have been widely established. We simulated a CAT in the calibration sample, using the answers on paper and pencil questionnaires as if they were given in a CAT and calculated the resulting person locations (CAT scores). We assume that in community-based PCH about 30 items is the maximum number feasible, and limited the number of items to be used in this CAT to 30. We used Fisher’s information Index for the selection of the next item in the CAT. Using the scores from this simulation we did a Receiver Operating Characteristics analysis with a clinical CBCL TPS as criterion and chose that point that resulted in a specificity of 0.90 as cut-off point.

The exact estimate of the person locations, however, will vary somewhat with the number of items used in the CAT. In order to assess the effect of this variation we repeated the analyses with a fixed number of 5, 10 and 20 items and also with no limit to the number of items, but continuing until the person locations had been estimated with 95% accuracy. In all these CATs the first item was chosen at random. We calculated the sensitivities and specificities for all these analyses and inspected the differences, in order to verify that the maximum of 30 items we used was a sensible one.

Finally, we evaluated the validity and efficiency of the CAT. The validity was assessed by means of a simulated CAT in the independent validation sample. In this simulation we aimed to assess, with an accuracy of 95%, whether a person scored above or below the chosen cut-off point. In other words, the CAT was stopped when the 95% Confidence Interval of the estimated person location did not overlap anymore with the chosen cut-off point. This procedure is known as clinical decision adaptive testing. Again, the starting item was chosen at random and Fisher’s Information index was used to select the next best item. We assessed the validity of the estimated person locations by calculating the Area Under Curve (AUC), sensitivity and specificity with a clinical CBCL TPS and current treatment status as criteria.

The efficiency of the procedure was evaluated by calculating the number of items needed in this simulated CAT and the number of respondents for whom the CAT resulted in 95% certainty on a score below or above the chosen cut-off point.

For the IRT analyses we used the RUMM 2020 software. All other analyses were done using SPSS.
8.3 Results

8.3.1 Are the items on emotional and behavioral problems suitable for an IRT-based CAT?

Of the 205 items in the four questionnaires 190 met the criteria for a CAT: they had an outfit of less than 1.7. Most items that had to be removed came from the CBCL (13 out of 15). The Person Separation Index was 0.93, indicating a high reliability. The DIF analyses showed that almost all estimates were not modified by gender and ethnicity. Only 8 of the 190 items showed some DIF; five items in relation to gender (sexual problems, running away, attacking others, being ill without physical cause and problems with teachers) and three in relation to ethnicity (tantrums, not being assertive, talking about suicide). Most of these problems have a very low prevalence and will therefore have a small overall impact on the final estimations. We therefore decided not to remove these items.

Figure 1 presents the estimated item locations calculated for the remaining items and split by questionnaire. As mentioned before, these item locations are indications of the level of severity. The most severe items on the right (concerning very serious problems) were items from the CBCL, which in general appeared to have more severe items than the other three questionnaires.

Figure 1 Estimated locations of the items in the four questionnaires in the calibration sample
8.3.2 Determining the cut-off point

After the item locations had been estimated, we did a CAT simulation on the calibration sample with a fixed number of 30 items. Figure 2 presents the number of respondents by the calculated person location on the latent scale, by CBCL TPS, divided into normal, borderline or clinical. The ROC analysis showed that with a cut-off point of -1.9 the specified specificity of 0.90 was reached. The sensitivity for a clinical CBCL TPS at that point was 93%.

Table 1 presents the effects in terms of AUC, sensitivity and the specificity indices in relation to the use of different numbers of items in the CAT. The specificity shows little variation; using a fixed number of 5 or 10 items results in a decreased sensitivity. The results for a CAT with 20 or 30 items and for a CAT that continues until the 95% Confidence Interval does not overlap with the cut-off point, are very similar.

![Figure 2: Distribution of estimated person locations in the calibration sample by CBCL classification](image)

<table>
<thead>
<tr>
<th>Criteria to stop the CAT</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Area Under Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. no of items: 30</td>
<td>93%</td>
<td>90%</td>
<td>0.97 (0.96 - 0.98)</td>
</tr>
<tr>
<td>Max. no of items: 20</td>
<td>90%</td>
<td>89%</td>
<td>0.96 (0.95 - 0.97)</td>
</tr>
<tr>
<td>Max. no of items: 10</td>
<td>79%</td>
<td>88%</td>
<td>0.92 (0.90 - 0.94)</td>
</tr>
<tr>
<td>Max. no of items: 5</td>
<td>54%</td>
<td>88%</td>
<td>0.90 (0.84 - 0.89)</td>
</tr>
<tr>
<td>Estimation of person location with 95% accuracy</td>
<td>92%</td>
<td>90%</td>
<td>0.97 (0.96 - 0.98)</td>
</tr>
</tbody>
</table>
8.3.3 **Validity and efficiency**

In the validation sample, the ROC analyses showed that the CAT did very well in the identification of children with a clinical TPS; the AUC was 0.92 (CI: 0.85 – 0.99). With the chosen cut-off point sensitivity was 0.89 (CI: 0.71 – 0.97), with a specificity of 0.91 (CI: 0.87 – 0.93). Kappa was 0.53.

Using treatment status as criterion the AUC was 0.74 (CI: 0.63 – 0.84). The sensitivity for current treatment status was 0.55 (CI: 0.37 – 0.72), with a specificity of 0.89 (CI: 0.85 – 0.92). Kappa was 0.32.

Overall, in relation to the CBCL TPS, the CAT selection procedure resulted in a correct classification of 91% of all children involved. The CAT resulted in a correct classification for the large majority of cases with normal (96%) or clinical scores (89%). However, 20 (77%) of the 26 cases with a score in the CBCL borderline range, had an elevated CAT score.

Figure 3 presents the number of items needed to reach convergence, i.e. to assess with 95% certainty whether the respondents had a true score below or above the chosen cut-off point of -1.9. In 40 cases (10%) convergence was not possible with less than 100 items. They had a mean person location of -1.88 (sd=0.18); i.e. very near the chosen cut-off point. Their mean CBCL TPS was 28.4 (sd=7.1); 25% of them had a CBCL TPS in the borderline range; 5 % in the clinical range.

For the other 351 cases, the mean number of items used was 11.5 (sd=13.0). For 37% of the respondents the procedure converged with less than 5 items; for 57% up to 9 items were needed. For 74% up to 20 items were used and for 82% up to 30 items.

*Figure 3  Efficiency of the CAT procedure: percentage of persons for whom a score above or below the cut-off point could be estimated with 95% accuracy, by number of items used to achieve convergence.*
The mean CBCL TPS for respondents for whom less than 5 items were used in the CAT was 10.8 (sd=10.5). We checked the convergence between the CBCL TPS based classification and the CAT classification for these respondents. In 98% of the cases the classification was identical. The CAT resulted in a below cut-off point score for 2 respondents with a clinical CBCL TPS and one respondent got a CAT score above the cut-off point with a CBCL in the normal range.

8.4 Conclusion

This study showed that IRT-based Computerized Adaptive Testing indeed resulted in an accurate, yet very efficient identification of children with psychosocial problems. Most of the items of the four questionnaires investigated met the requirements of an IRT model, needed to incorporate them in a CAT. A simulation study showed that the procedure identified children with a clinical CBCL TPS with high sensitivity and specificity. For 90% of all cases we could determine with 95% certainty whether they had an elevated score. In order to achieve these results, on average only 11.5 items were needed. For more than half of the children less than ten items were needed.

There are, of course, other, more traditional techniques for reducing test length. However, other than more traditional approaches, an IRT-based CAT provides high measurement quality, by adjusting items used in the assessment to the individual being tested. This has the additional advantage that this individual is not being confronted with items that are not relevant in his situation and that possibly may be shocking for him or her. The inclusion of items of the SDQ, PSC and SDQ in the item pool used for this CAT offers therefore the advantage that more items are available that are suitable for parents of children with no or few problems.

Fit with the literature

Our finding that an IRT-based CAT can result in accurate assessments with far less items than tests based on traditional psychometrics is fully comparable to findings in other studies, applying IRT CAT techniques in the fields of intelligence and school achievement assessment,15,21 and in the field of Quality of Life.17,40 The first studies on the application of IRT models in the field of the identification of psychosocial problems have now been published,18-20 and these studies came to similar conclusions. Compared to other studies, known to us and regarding CAT and mental health, our study and the study by Gardner19 are the only ones to focus on a rather broad concept, rather than on more specific problems, like Gardner18 and Fliege et al.20 Gardner19 used the PSC as criterion. As we used the more widely validated CBCL as one criterion, our study is a stronger argument for the usefulness validity of CAT-based procedures in the field of mental health.

Gardner19 evaluated the extent to which a multidimensional adaptive test could be used to replicate screening decisions based on the Pediatric Symptom Checklist. He found a very high correspondence between the Adaptive PSC and the original 35 items PSC (kappa=0.84), higher than the corresponding figure we found. The mean number of items he needed to achieve this was 12 items, out of 35, whereas we needed a mean of
12 items, to replicate the screening decision based on the 120 item CBCL. It is not exactly clear why he found a higher correspondence than we did. Our cut-off point was not chosen in order to maximize kappa, but had we done so, our kappa would still be lower than Gardner’s. One explanation may be that Gardner used a multidimensional model, whereas we used the simple one-dimensional Partial Credit Model. Another explanation might be that Gardner limited himself to PSC items, whereas we used items from four questionnaires. Thus, in our study there is less overlap between the items in the CAT and the criterion measure. This is probably the main reason that Gardner’s study resulted in a higher kappa.

Strengths and Limitations
This study has several strengths but also limitations. A major strength is that it concerns a community-based sample of children with high response rates and representative for the population under care. Moreover, the pool of items used was much broader than the criterion that we used, leading to a relative independence of the criterion from the items used for prediction. This was reinforced by the use of separate samples for the construction of the CAT and for its validation. However, our study also has several limitations. Though more independent than in previous studies, some of the predicting items will also be included in the criterion. This is a limited problem, though, if one tries to reach the best short alternative for a longer questionnaire, as we did. Moreover, we simulated a CAT based on answers given to a full questionnaire, which is a deviation from the real practice set-up. A next stage will certainly be to evaluate the CAT in a real-life situation. Finally, although we had a rather large sample, our validation sample was relatively small, implying the need for a large scale replication. Anyhow, our study provides a valid assessment of the potentials of an IRT-based CAT for PCH practice.

Implications
The most important implication of our study is that IRT-based CAT appears to be a very feasible and promising tool to improve the identification of psychosocial problems in PCH. As such it earns a quick passing through to the daily practice of well-child care and maybe even of pediatric care in general. Before having a final pass to clinical practice, several aspects have to be studied more thoroughly, though. This in particular concerns the use of our simulated version in a real-life situation, with parents filling out the CAT on real computers. Currently, a beta-version for this aim is available at www.uwkind.nl, but this is Dutch only and protected by passwords and firewalls to preserve patient-confidentiality. A formal assessment of this implementation in daily practice is the next step for research. Similarly, our findings have to be replicated in other settings and maybe using other item pools as well. Anyhow, this new technology may provide a push to improve the quality of the identification of psychosocial problems in PCH.
References


Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000 September;38(9 Suppl):II73-II82.