Can Western developmental screening tools be modified for use in a rural Malawian setting?

M J Gladstone,1 G A Lancaster,2 A P Jones,3 K Maleta,4 E Mtitimila,5 P Ashorn,5 R L Smyth6

ABSTRACT

Objective: To create a more culturally relevant developmental assessment tool for use in children in rural Africa. Design: Through focus groups, piloting work and validation, a more culturally appropriate developmental tool, based on the style of the Denver II, was created. Age standardised norms were estimated using 1130 normal children aged 0–6 years from a rural setting in Malawi. The performance of each item in the tool was examined through goodness of fit on logistic regression, reliability and interpretability at a consensus meeting. The instrument was revised with removal of items performing poorly.

Results: An assessment tool with 138 items was created. Face, content and respondent validity was demonstrated. At the consensus meeting, 97% (33/34) of gross motor items were retained in comparison to 51% (18/35) of social items, and 86% (69/80) of items from the Denver II or Denver Developmental Screening Test (DDST) were retained in comparison to 69% (32/46) of the newly created items, many of these having poor reliability and goodness of fit. Gender had an effect on 23% (8/35) of the social items, which were removed. Items not attained by 6 years came entirely from the Denver II fine motor section (4/34). Overall, 110 of the 138 items (80%) were retained in the revised instrument with some items needing further modification.

Conclusions: When creating developmental tools for a rural African setting, many items from Western tools can be adapted. The gross motor domain is more culturally adaptable, whereas social development is difficult to adapt and is culturally specific.

Eighty percent of the world’s disabled population live in low income countries, many of these in Africa.1 The World Health Organization has made early identification of children with disabilities a high priority, especially as early rehabilitation may reduce the impact of impairments.2–3 To identify these children and provide basic services, developmental milestones need to be clearly identified. Furthermore, clinical studies investigating interventions in children require normal parameters.

When child development is assessed in clinical studies in developing countries, Western developmental tools are often utilised.4–5 These include the Bayley scales,6 the Griffith’s,7 the McCarthy scales8 and the Denver II,9 all designed and validated in Western countries. These tools may be tailored for use in non-Western settings. Often translation (changing of the language used) is all that is carried out.10–11 If this is not accompanied by a process of adaptation, translation alone may not allow completely for local expressions and customs, therefore leading to misinterpretation of results.12 In other settings, tools are adapted and items are modified and in some cases new items are created for use within a Western tool.13 Sometimes these tools are piloted (tried out before use)14 and validated (assessed that they are measuring what they are supposed to be measuring) in the local population.15 Even these adapted tools, however, are of limited value without normal ranges for their defined population. Standardisation studies (finding norms for a population) have taken place in many non-Western countries mainly using the Denver Developmental Screening Test (DDST) in a translated and occasionally adapted form.16–18 but none of these studies was in Africa. Only two studies have attempted standardisation in Africa, one using a translated form of the Bayley scales with an urban black South African population19 and the other on a limited age range in a rural Nigerian population.20

It is clear that Western developmental assessment tools may include tasks and materials which are completely alien to other cultures. These tools may therefore fail to identify and assess children adequately in cultural settings other than those for which they were created.21 This may be less of a problem when comparing groups of children, but when Western tools are used alone as an outcome measure, culture may have an effect. In theoretical studies, culture has been demonstrated to have an influence on child development, particularly in the area of social development.22–24 Cognitive abilities such as memory, categorisation techniques and pattern recognition have also been reported to be influenced by culture.25–28 Even gross motor development may possibly be affected by culture.29–31

In this study, we aimed to create a simple, culturally appropriate developmental assessment tool adapted and modified from Western tools and standardised for use in rural Malawi. The first stage in the development of this tool was to identify which items from Western tools (eg, the DDST or Denver II) were not relevant to the age-appropriate experiences of rural Malawian children. These items were then replaced with ones more appropriate to this cultural context. We did this firstly by holding focus groups to agree which items should be replaced and to create alternative items. All items (both retained and new) were then validated and standardised in a large population study. The performance of all items was examined in a consensus meeting and a revised instrument proposed.
METHODS
Setting and study population
This study was a substudy of the Lungwena Child Survival Study (LCSS), a prospective family cohort study looking at gestational health and the growth, development, morbidity and mortality of rural Malawian infants and children. Lungwena is an area in southern Malawi where a government health centre serves an approximately 100 km² rural area with some 17 000 people in 23 villages. Most of the inhabitants are Muslims of the Yao tribe. The literacy rate is low and subsistence farming and fishing are the main occupations. The original cohort for the LCSS was enrolled between June 1995 and August 1996. All pregnant women presenting for antenatal care were eligible for the LCSS and 97% of the population of pregnant women in the area, at that time, were enrolled in the study. Details of recruitment, collection of background data and follow-up have been described previously.32 33

The population of children used for this study is the original LCSS cohort of children aged 3.5–6.5 years and younger siblings aged 0–3.5 years. Out of the 1237 LCSS children and siblings available, 1197 were seen, with 40 families either refusing to take part or not being available. The ages of the children were known from LCSS birth data or from the “health passport” given to mothers at the birth of their baby where the date of birth is recorded and which almost all mothers carry with them for all health appointments. A quota sampling strategy was used as in the DDST and Denver II34 with target numbers of children being sought in each of 33 age groups (see supplementary table A). A total of 67 children were excluded due to premature birth (34 weeks or less measured by fundal height at the antenatal clinic),32 twin birth or significant disability including severe malnutrition (weight for height z score of less than −2), leaving 1130 children in the final analysis.

The LCSS received approval from the National Health Science Research Committee in Malawi (HSRC 95/94). Informed verbal consent was sought from each mother at the beginning of the LCSS and again before a development assessment was carried out.

Creation of the developmental assessment tool
The Denver II, DDST and Griffith’s instruments were examined by the Malawian research team. Items considered to be culturally appropriate were included and translated, whereas those considered inappropriate (such as “prepares cereal” or “plays board/ card games”) were removed. New items and modifications to

Figure 1  The process of creating a more culturally appropriate developmental assessment tool.

Examination of Western tools - Denver II (125 items), DDST (105 items) and some items from Griffith’s

Translation of items

4 Focus groups (2 days each) for creation of new items and modification of items

119 items in revised assessment tool

Face, content validity
Six items added
1st piloting - 20 children

125 item instrument

Feedback and training for research team

Modifications (re-translation and adaptation)
13 items added
2nd piloting - 20 children

138 item instrument

Feedback and training for research team

Standardisation on 1130 children using logistic regression and goodness of fit

Interobserver (7–14 days) on 46 children and intra-observer (7–14 days) on 25 children

Respondent validation

Concensus meetings

Revised instrument with 110 items

Original article

Examination of Western tools - Denver II (125 items), DDST (105 items) and some items from Griffith’s

Translation of items

4 Focus groups (2 days each) for creation of new items and modification of items

119 items in revised assessment tool

Face, content validity
Six items added
1st piloting - 20 children

125 item instrument

Feedback and training for research team

Modifications (re-translation and adaptation)
13 items added
2nd piloting - 20 children

138 item instrument

Feedback and training for research team

Standardisation on 1130 children using logistic regression and goodness of fit

Interobserver (7–14 days) on 46 children and intra-observer (7–14 days) on 25 children

Respondent validation

Concensus meetings

Revised instrument with 110 items

Figure 1  The process of creating a more culturally appropriate developmental assessment tool.
Western test items were then created through discussions with a series of focus groups. Key informants were the eight local research workers. They were all women of child-bearing age with at least 8 year’s education and research experience of at least 5 years. Themes relating to developmental milestones were discussed and ideas from these sessions were used to create new items. Illustrations were made for most items in the instrument and used as prompts for the research workers. Some came with permission from Disabled village children. 

Face validity and content validity were assessed by all research assistants, five Malawian paediatricians, a language expert from the University of Malawi and six medical students at the College of Medicine, Malawi. Once the new instrument was created, the team was trained in its use and it was piloted in two stages. At each stage, feedback and training were given and problematic items were re-adapted or re-translated. The process of creating and refining the more culturally appropriate tool is shown in fig 1.

### Standardisation using a normal population sample

Overall, 1197 children were assessed on one occasion between February 2000 and April 2001 on a home visit by research assistants. The assessment took approximately 35 min to complete and where possible, items were directly observed. In a few cases a report was given, for example “does he go to the toilet by himself?” Items were scored as either pass or fail, or “don’t know” if the child was uncooperative or unwell. Items were asked until the child failed seven items in a row. Data entry and analysis were carried out using Microsoft Excel 6.0, SPSS 11.1, Stats-direct and STATA computer programs. Each child in the study was identified by a code. Data were checked prior to analysis and any outlying results were reviewed.

Standardisation is the process of determining normal age ranges for which children pass the items for a developmental assessment tool. A logistic regression analysis was carried out with decimal age and sex as explanatory variables. The observed and predicted probability of passing was determined and graphs were drawn for each item. The goodness of fit of the graph was visually assessed and discrepancies reviewed. To determine statistically whether or not the fitted curve was a sufficiently good representation of the data, a goodness of fit statistic was calculated. If this was significant at the 5% level, indicating a poor fit, then the data were re-examined and refitting was done.

---

**Table 1** Examples of specific items added or removed during the process of creating new more culturally appropriate tool

<table>
<thead>
<tr>
<th>Domain of development</th>
<th>DDST or Denver II items removed at stage 1</th>
<th>New items replaced at stage 1</th>
<th>Denver II/Griffith’s items removed or needing modification after consensus meeting</th>
<th>New items removed or needing modification after consensus meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gross motor</strong></td>
<td>1. Head up to 45° 2. Walks up steps 3. Pedals tricycle</td>
<td>1. Dances to music 2. Kneels well to be respectful 3. Climbs and gets down 4. Carries object on head with one hand 5. Pushes wheel with stick 6. Carries water on head – no hands</td>
<td>1. Pulls self to stand (not understood by research assistants) 2. Touches toes with legs straight (difficult to teach)</td>
<td>1. Dances (not specific enough) 2. Carries object on head with one hand (need to specify exact object) 3. Carries water on head with no hands (need to specify size)</td>
</tr>
</tbody>
</table>
using triple split spline regression. The ages corresponding to the 35th and 65th percentiles were calculated from the original fit to determine the cut-points. For some items that performed less well, the cut-points were chosen by viewing the graphs to facilitate a good fit. Three logistic curves were then fitted, one for each region, based on the split. Any items with significant gender effects were removed or considered for further modification to ensure the tool was applicable to all children irrespective of gender. Using the predicted probabilities found from the logistic regression analyses, the ages corresponding to 25%, 50%, 75% and 90% of the children passing were determined for each item. These were then used to plot the age norms of achievement of each milestone in a box-type representation.

Reliability of the items
Reliability for each item was tested by using two subsamples of 60 (inter-observer) and 28 (intra-observer) randomly selected children who were seen at 7 and 14 days after initial assessment. Of the 60 children, 46 completed the follow-up using two different examiners (inter-observer), while 25 of the 28 children used the same examiner (intra-observer). All items in the tool were assessed for both types of reliability. Kappa statistics (κ) with 95% confidence intervals (CI) were used to calculate the degree of observer agreement for each question. Positive values of 0 to <0.2 indicate poor agreement, >0.2 to 0.4 fair agreement, >0.4 to 0.6 moderate agreement, >0.6 to 0.8 good agreement and >0.8 to 1 very good agreement.

Respondent validation was carried out after the preliminary analysis. This method of validation involves the reporting of findings back to the participants. Findings were fed back at the end of the study to the Lungwena Health Centre Management Committee. This consisted of four chiefs, one overall representative and three women representatives, all from the local area.

Consensus meeting
Once all the items were analysed, an expert panel (MG, AJ, EM and GL), which included a Malawian paediatrician, met to review the results and decide which items should remain, which should be modified and which should be removed. Items were judged on their graphical representation, and goodness of fit on logistic regression, reliability and subjective ratings of “interpretability” by participants and researchers.

RESULTS
A tool with 138 items (34 gross motor (GM), 34 fine motor (FM), 35 language and 35 social items) was created. An example of the tool is shown in supplemental fig B (see supplementary data). Most (58%) items were from the DDST and Denver II, with a small percentage (9%) from the Griffith’s instrument. Many items in the GM (82%), language (77%) and FM (70%) sections were directly translated from Western tests with modifications mainly in the FM section. Only 37% of the social items were taken from Western tests. The first two columns of table 1 provide examples of many of the items that were removed from the DDST or Denver II, and shows the newly created items which replaced them.

The face validity and content validity of the tool were tested. The modified instrument appeared to those questioned to cover development in children in ways that were important, and it was judged to examine in a fully comprehensive and logical fashion the domains of child development for children in Malawi. It was therefore considered to have good face validity and content validity. Most items were found to be acceptable for studying children’s development in this setting through respondent validity. The pictures as prompts were found to be particularly helpful to the researchers in the field.

Examples of graphs created through logistic regression during the standardisation procedure and where triple split joined
regression was used, are shown in fig 2. In terms of goodness of fit on logistic regression and on spline regression, social items had the highest number of poor fits (51%, 18/35), sex being an independent predictor in some of these (23%, 8/35) (see table 2). A larger proportion of the newly created items had a poor fit on logistic regression (15%, 7/46) and an effect of sex (17%, 8/46) than those from the Western tools. The few items not attained by 6 years came from the fine motor area of development and included “draws a man with 6 parts” and “draws a square”. The results of the Lungwena milestones for the language section of development are shown in fig 3. The other areas of development are described in supplemental fig C (see supplementary data).

Reliability results are also shown in table 2. For inter-observer reliability, 82% (113/138) of the questions had moderate to very good reliability (κ>0.4). There are no figures in the Denver technical manual for inter-observer reliability for comparison. Intra-observer reliability demonstrated moderate to very good reliability (κ=0.4) for 75% (106/138) of the questions. This compares well with Denver II figures, where 81% of their items had a κ>0.4. In relation to the domains of development, GM items had the best overall inter-observer (29/34 items) and intra-observer (32/34 items) reliability with κ=0.4. Items from the social area performed less well, with only 74% (26/35) of the items on inter-observer

Table 2 Decisions regarding suitability of items within each domain of development and the source of the question

<table>
<thead>
<tr>
<th>Domain of development</th>
<th>Source of question</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Denver or DDST</td>
</tr>
<tr>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>27</td>
</tr>
</tbody>
</table>

DDST, Denver Developmental Screening Test; FM, fine motor; GM, gross motor; Lang, language.
reliability and 60% (21/35) of the intra-observer items having a $\kappa>0.4$. In relation to the source of the item, more of the locally-derived items had poor inter-observer (33/46) and intra-observer (15/46) reliability ($\kappa<0.4$) in comparison to those items derived from the Denver II (12/80 and 8/80).

After a consensus meeting, 110 of the 138 items (80%) were retained in the revised instrument, with some needing further modification. Only 69% (52/46) of the newly created items were retained in comparison to 86% (69/80) of the DDST or Denver items used (see table 2). The results of this meeting giving examples of items removed are detailed in the last two columns of table 1.

**DISCUSSION**

We have demonstrated that many items from Western tools can work well when adapted and translated for other settings. They have already had their own rigorous reliability and validity studies carried out in the West and therefore are more likely to be robust in use. However, through our focus group, validation and piloting work, we have also demonstrated that in all domains of Western tests (such as the DDST), there are some items which are culturally inappropriate for a rural African population. For example, questions such as “prepares cereal” or ‘plays board games/card games” are uncommon activities for children in rural Africa. Also, the pink doll in the DDST kit was terrifying to most children when used in piloting; many children had never seen anything like it and many screamed. It would have been unlikely that we would have been able to get them to sit down and “feed the doll”. Some of the naming questions in the Language section of the DDST or Denver II have pictures of objects that children, at least in the part of rural Africa studied, have never seen before, such as a horse and a car. This makes it difficult for them to name them, especially as many children have also never seen a book at their age, or pictorial representations of many objects.

In the creation of new items, however, many newly created items were less reliable, more sex-specific and had poorer goodness of fit in logistic regression. This was most evident in the social domain and least evident for gross motor skills. Social skills seem to have the least “universality” and in measuring them, we need to question the appropriateness of the concepts being measured in such different settings. When measuring “social skills” we may be determining the ability of the child to have learned important skills instilled by parents and carers in particular cultural settings, but this can only be measured if pertinent skills are tested for. The difficulty when creating new social items for a tool such as this, is that the items must be specific enough to distinguish between the developmental age ranges of children, but also be clear and easy to explain in a developmental tool. This will continue to be a challenge.

It was not a primary aim of our study to compare our results with the Denver II or DDST. A formal statistical comparison has not been possible; however, when comparing our charts with those from the Denver II or DDST on gross comparison, it does seem that there are obvious differences in milestones with children from the West. For example, the item “combines two words” in the Denver II is attained at between 17 and 21 months, whereas in our sample this was obtained at between 21 months and 2 years 4 months. This demonstrates the importance and necessity of creating norms for a given African population, as they are likely to be different from those in the West.

A second phase of work is currently underway using the methodology that we have formulated in this first study to refine a further tool with a larger standardisation sample. This work will include creating a scoring system, and carrying out more detailed reliability measurements and further validity tests of between-group and construct validity. Once this new version has been created and has undergone the strict procedures that we have instituted in our methodology, we hope to have created a tool that may benefit community health workers in other rural settings in Africa after local validation. The complete tool may also be used by research workers who are investigating developmental outcomes as part of their intervention strategies.

**What is already known on this subject**

- The ages of attainment of developmental milestones can differ with the cultural background of the child, although there are very few data from Africa.
- Developmental assessment in African settings is often carried out using translated and adapted Western tools, with a dearth of tools adapted, validated and standardised specifically for rural African children.

**What this study adds**

- A procedure using validity, reliability, goodness of fit on logistic regression and expert consensus has been devised for assessing newly created questions or items adapted or taken from Western developmental assessment tools.
- Many items from Western developmental assessment tools can work in a rural African setting. Gross motor items are most reliable, whereas items within the social area of development need to account for cultural differences.

**REFERENCES**