STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2272–2283
Published online 5 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2351

An interval scale for development of children aged 0–2 years

Gert Jacobusse[*,†], Stef van Buuren and Paul H. Verkerk

TNO Prevention and Health, Leiden, The Netherlands

SUMMARY

Measurement of development is often less precise than that of height and weight. Developmental scores are typically based on passing one or more developmental markers, but do not have interval scale so calculating differences between scores can be nonsensical. Age-specific standardized scores are sometimes used, but fail to have a common metric that allows comparison of developmental scores across age. The goal of this study is to develop a quantitative developmental score (*D*-score) with improved measurement characteristics. The basic assumption of the *D*-score is the existence of a common continuous scale for the development. Scores of 2151 children between 0 and 2 years on a Dutch developmental instrument were analysed. Application of the Rasch Model resulted in excellent reliability and satisfactory fit. This indicates that the new quantitative *D*-score succeeds in representing outcomes of the instrument on a common interval scale. Age-conditional reference values for the *D*-score were derived by means of the LMS method. The definition of the *D*-scores is not specific to age, so the *D*-score of a measured person can be compared to the *D*-score of another person of a different age. Difference scores between sessions can be used to evaluate developmental velocity on the individual level. To our knowledge this is the first developmental scale for children with such properties. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: measurement; child development; Rasch model; LMS model; developmental monitoring

1. INTRODUCTION

Growth and development during the first years of life are the basis of lifelong outcomes. It is important to monitor these processes, in order to signal and treat any delay as soon as possible. The monitoring of growth is a well established practice. Measurements of growth like height and weight are very precise, and growth diagrams and referral criteria have been developed. On the other hand, instruments for monitoring development are less precise, and only roughly indicate the development of a child. Measurements often consist of a pass or

[*] Correspondence to: Gert Jacobusse, Department of Statistics, TNO Prevention and Health, P.O. Box 2215, 2301 CE Leiden, The Netherlands.
[†] E-mail: gw.jacobusse@pg.tno.nl

Received 16 December 2004
Accepted 11 July 2005

fail outcome on a set of dichotomous items. The resulting scores are not expressed on a convenient quantitative scale like centimeters or kilograms.

The Bayley scales of infant development are a well known example of an instrument to measure development. These scales were originally proposed in the 1950s. Since then several efforts have been made to improve its measurement properties [1]. Outcomes on subscales of this instrument are summarized by a standardized score. The development of a child is expressed as a relative position within the distribution of all healthy children of the same age, assuming that outcomes of healthy children are normally distributed. Three main shortcomings accompany the currently used scores:

1. The measurement scale is relative to a specific population, the 'norm group'.
2. There is no common metric to compare outcomes. Difference scores are not meaningful because there is no underlying quantitative scale.
3. The exact meaning of the same score may differ across age; it is not possible to quantify a child's progress in time in terms of a gain in developmental units.

In order to deal with these weaknesses, we will express outcomes of the instrument on an interval scale. This scale provides a common metric so that difference scores are meaningful, even across age and across different sets of items. Scores must be independent of the population that is being measured. The Van Wiechen Scheme [2] is the Dutch equivalent of the Bayley scales. In this article we suggest to improve its measurement properties by applying a method to express developmental scores on a quantitative scale. We call the quantitative developmental scores '*D*-scores'. The *D*-scores allow for sensible comparisons both within and across age.

1.1. The Van Wiechen scheme

The Van Wiechen scheme was developed during the sixties by Dr H. J. van Wiechen, a Dutch family doctor. He originally applied the scheme for early detection of spasticity in young children. Measurement and diagnostic properties of the scheme have been investigated and the scheme has been further developed by Schlesinger-Was [3]. Nowadays, Child Health Care Centers in The Netherlands routinely use this scheme to monitor the development of children from birth to four years of age [4].

The scheme consists of a set of 57 developmental indicators. The youth health care physician assigns a pass or fail score to each indicator for a given child. Indicators are divided into different sets, each targeted at children of a certain age. Table I contains a description of the items.

An age-appropriate set of indicators is administered during sequential sessions to follow a child in time. Indicators that are meant to be administered at a given age are chosen so that the majority of children (about 90 per cent) will pass. This enables the monitoring of usual development and the detection of possible deviations from it. A fail score can be a signal of a delayed development, although one fail score is usually not enough reason to immediately further investigate the child. The Van Wiechen scheme is more used as a monitoring than as a screening instrument, and referral decisions heavily depend on the individual assessment of the physician. Clear criteria for signalling a delayed development have not been formulated.

Currently, the outcome on each indicator in the Van Wiechen scheme is interpreted separately. This has the advantage that a fail score can be located to a very specific area

Table I. Item formulations of the Van Wiechen scheme for development between 0 and 2 years, split according to target age.

Set (Target age)	Item formulation
1 (4 weeks)	Fixates eyes Reacts to speech Moves both arms equally as much Moves both legs equally as much Lifts chin
2 (8 weeks)	Smiles back Follows with eyes and head
3 (13 weeks)	Hands open now and then Looks at own hands Vocalizes responsively Remains positioned when lifted under the armpits Holds head up 45° in prone position
4 (26 weeks)	Hands playing in midline Grasps toy within reach No head lag when pulled to sitting position Turns head to sound When lifted vertically, legs bended or trampling Holds head up 90° in prone position
5 (39 weeks)	Transfers toy easily, hand to hand Picks up one small toy, then second Plays with both feet Rolls from prone to supine and back Holds head up in sitting position Sits with stretched legs Says ‘dada’, ‘baba’ or ‘gaga’
6 (52 weeks)	Sits without support Picks up crumb between thumb and index finger Crawls Pulls himself to standing position Waves ‘bye bye’ Jabbering
7 (15 months)	Gets cube into and out of box Plays ‘give and take’ Crawls, with belly lifted off the ground Walks while holding furniture Understands some simple words Uses two words
8 (18 months)	Makes tower of two cubes Explores room Uses three words Identifies two named objects Walks on its own Throws ball without falling

Table I. *Continued.*

Set (Target age)	Item formulation
9 (24 months)	Makes tower of three cubes Imitates everyday activities Drinks from cup Makes 2-word sentences Puts ball in box when asked Squats Walks well
10 (30 months)	Makes tower of six cubes Puts round figure into place Takes off a cloth (shoe, sock, trousers) Eats with spoon without help Calls itself by name or 'I' Identifies pictures in book Kicks ball away

of development. However, there are some limitations to this approach. The measurement error of an outcome on one single indicator is quite large. Furthermore, the interpretation of outcomes involves multiple testing, and thus an inflated type I error rate. A more reliable estimate can be attained by a composite score that is based on multiple indicators, as in the Bayley scales. Our strategy is to improve the measurement properties of the Van Wiechen scheme, by combining outcomes on all indicators into one score. In contrast to the Bayley scales, scores will be expressed on a common quantitative scale.

1.2. Towards quantitative scores for development

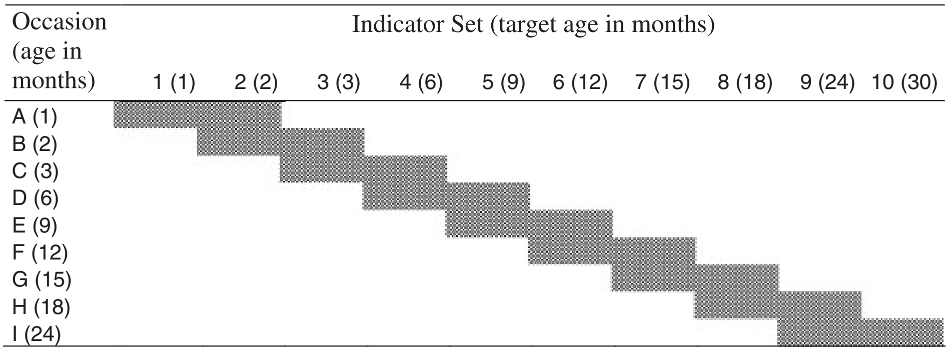
In this article we presume that a common underlying scale for the development of children from birth to two years of age exists. *D*-scores cannot be measured directly, but will be estimated from the outcomes of the Van Wiechen scheme indicators. We apply a model to estimate the scores, and examine the fit of that model to test whether the common scale for development succeeds in representing the outcomes of the Van Wiechen scheme.

The method to estimate interval scale *D*-scores is outlined in the next section. In the results section, we illustrate the opportunities that the new *D*-scores create. Analogous to growth diagrams, a chart that quantifies the relationship between age and development will be created. Potential applications and pitfalls will be discussed.

2. METHOD

2.1. Data

We analysed longitudinal data collected within the 'Social medical survey of children attending child health clinics' (SMOCC) project [5]. Subjects were all 2151 children born between April 1988 and October 1989 attending one of 21 Child Health Care Centers in The



Occasion (age in months)	Indicator Set (target age in months)									
	1 (1)	2 (2)	3 (3)	4 (6)	5 (9)	6 (12)	7 (15)	8 (18)	9 (24)	10 (30)
A (1)	▨	▨								
B (2)		▨	▨							
C (3)			▨	▨						
D (6)				▨	▨					
E (9)					▨	▨				
F (12)						▨	▨			
G (15)							▨	▨		
H (18)								▨	▨	
I (24)									▨	▨

Figure 1. Data collection design, illustrating which indicator sets (columns) are administered at which occasion.

Netherlands. Data were sampled at nine sessions between birth and 24 months of age. At each session, a physician assessed whether a child could perform a set of developmental behaviours and tasks appropriate for the child's age. In the SMOCC study, each set (except the first) was also administered at the session before the target age. Approximately half of the children would pass the item on this more difficult set for that age. Figure 1 illustrates how indicator sets were distributed across sessions.

Note that the sampling design connects all indicator sets to each other. A child always shares a subset of items with children—themselves or others—at a later and an earlier occasion, because each indicator set (except 1 and 10) was evaluated at two occasions. Indicator set 2, for example, was evaluated during session A and session B. The indicators of both sets 1 and 3 can be related to set 2, which in turn enables relating indicator sets 1 and 3, although sets 1 and 3 were never evaluated together in one session. Using appropriate statistical modelling, this allows relating all indicators to each other.

2.2. Model

We fitted the Rasch Model (RM) to the data [6]. The RM starts from the assumption that there is a continuous latent trait θ that governs the probability of passing each developmental item for a given child. The latent trait θ can be interpreted as a kind of summary developmental score. Under the RM, a child i ($i=1,\dots,n$) at age t (in days) has a position θ_{it} on the latent scale, a number indicating the child's ability, or maturation, at age t. If all is well, θ_{it} increases with t as the child matures with age.

Let X_{ijt} ($j = 1,\dots,m$) be the outcome of child i at age t on the jth developmental item, where X_{ijt} only takes on values 0 (child does not pass) or 1 (child passes). Under the RM, outcome X_{ijt} on item j is characterized by the position δ_j on the latent trait, loosely interpreted as the 'difficulty parameter' of the item. In the sequel, we assume that δ_j is independent of age for all j, i.e. the difficulty parameter of the item does not depend on age.

For given θ_{it} and δ_j, the RM describes the probability that child i passes item j at age t as

$$P(X_{ijt}=1|\theta_{it},\delta_j)= \exp(\theta_{it}-\delta_j)/(1+\exp(\theta_{it}-\delta_j)) \tag{1}$$

which corresponds to the simple logistic model based on the difference between θ_{it} and δ_j. The probability of failing the item is equal to $1 - P(X_{ijt}=1)$. Note that if $\theta_{it}=\delta_j$, then $P(X_{ijt}=1)=1-P(X_{ijt}=1)=0.50$. In this model, the exponential functions of all indicators are assumed to have equal slopes, so that their relative difficulty is the same at different levels of θ.

In the case where $j>1$, the RM expresses the response probability of the combined response vector as the product of the individual probabilities. Let j and k $(j \neq k)$ be any pair of two items. The probability of observing the combined response vector X_{ijt} and X_{ikt} is simply equal to the product of the item-specific probabilities, i.e.

$$P(X_{ijt} \cap X_{ikt}|\theta) = P(X_{ijt}|\theta)\,P(X_{ikt}|\theta) \qquad (2)$$

The generalization to $j>2$ will be obvious. Condition (2) is known as the principle of local stochastic independence, and it is an integral part of the Rasch model.

A unique property of the RM is that the θ- and δ-parameters are separable. This implies that the D-scores θ_{it} can be estimated irrespective of difficulty of the items (δ_j). Vice versa, the difficulty of the items can be determined irrespective of the ability level of the calibration sample. If the assumptions of the RM hold, then scores θ have interval scale level properties [7].

2.3. Estimation

In practice, both θ- and δ-parameters are unknown and must be estimated from the data. The overlap between sessions that was illustrated with Figure 1 is essential for identification of both parameter sets. Children of a different age are judged on different indicators, but the link between indicators can be exploited by the RM to estimate a model for all indicators simultaneously, and express the scores of all children on one underlying scale [8]. We use the RUMM 2020 program [9] to estimate both parameters set for the data in Figure 1. RUMM 2020 uses the pairwise conditional method using principal components.

The pairwise conditional method estimates the difference between difficulty parameters δ_j and δ_k of item j and k, by comparing the number of persons passing item j and failing item k to the number of persons failing item j and passing item k. Only the subset of persons that responded to both items j and k is involved in this comparison, which makes the pairwise conditional method well suited to deal with incomplete data designs as in Figure 1. All pairwise comparisons are combined into one pseudo-likelihood function that must be optimized [10]. Once the item parameters δ_j are estimated, the person parameters θ_{it} can be obtained by maximizing the probability of obtaining the observed data given equation (1). D-score estimates θ are rescaled to have mean 50 and standard deviation 10.

2.4. Model fit

Two diagnostics are used to evaluate the adequacy of the model. The RUMM software provides a fit residual statistic per item. When the RM is true, this measure follows a standard normal distribution. Values above 3.0 indicate that unexpected deviations from the model occur (i.e. the relation between θ and $P(X_{ijt})$ is weaker than under the RM), while values below -3.0 indicate the raw probabilities depend more strongly on θ than predicted by the RM. In general, values between -3 and 3 are considered satisfactory. The second measure is the 'outfit mean square' statistic [11]. This is the mean squared standardized residual, a measure of how severe deviations from the model are. For each item, we calculate its value from the RUMM residuals output. The expected value equals 1, values inside the range 0.5 –1.5 are 'productive for measurement' [12].

The RM also assumes i.d.d. observations. This assumption is true for persons, but violated for repeated measurements within persons. Since sample size will be over-estimated, this violation will affect (lower) the standard errors of the parameters estimates. We expect that the effect in the estimates themselves will be minimal however. In order to test this supposition, we will also analyse a subset of the data in which only one (randomly selected) session measurement of each child is included, and compare the result to the full analysis.

3. RESULTS

Development, like growth, is closely related to age. Figure 2 shows the percentage of pass scores by age for all 57 indicators. The percentage of pass scores increases with age for all indicators. Remember that indicators were evaluated at the appropriate age and during one earlier session. Data at the left part of each curve (lower age) are generally obtained at the earlier session, while the right part represents pass scores at the target age. As expected, the percentage of pass scores is about 90 per cent for children of the target age. Note that curves on the left of Figure 2 have a steeper slope than those on the right. The percentage of pass scores increases faster for younger children.

The RM that was applied to estimate underlying developmental scores fits the data quite well. The separation index—an estimate of reliability—is 0.989. RUMM fit residuals range between -8.24 and 4.66, outfit mean square statistics vary between 0.23 and 1.84. Their complete distributions are given in Figure 3.

The order of RUMM fit residuals and outfit mean square statistics has the same meaning. Their correlation is 0.77, indicating that both statistics agree on which items over- (low

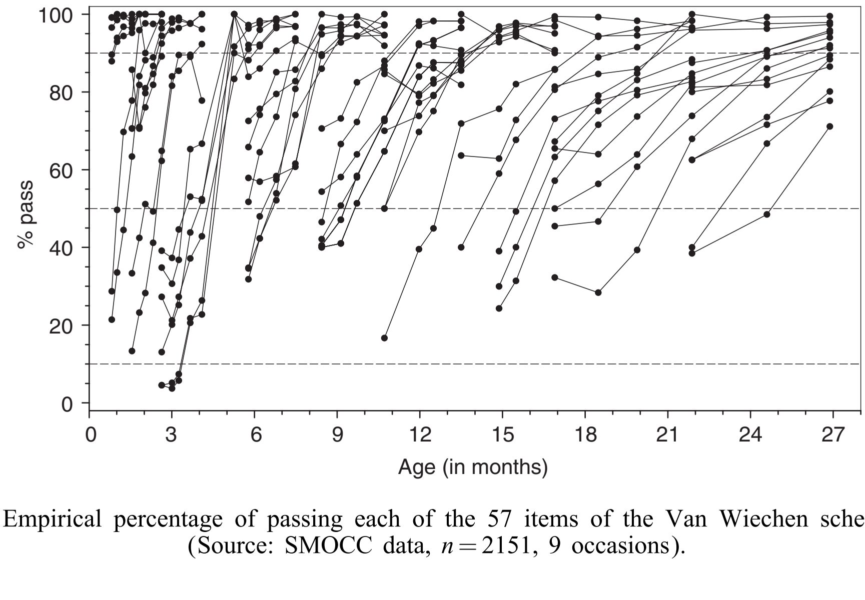


Figure 2. Empirical percentage of passing each of the 57 items of the Van Wiechen scheme by age (Source: SMOCC data, $n = 2151$, 9 occasions).

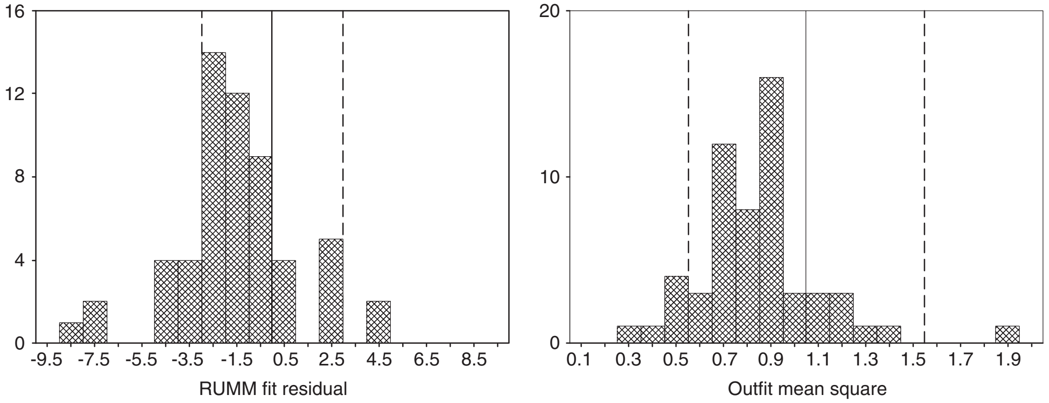


Figure 3. Distribution of the RUMM fit residual (left) and the Outfit mean square statistics (right) after fitting the Rasch Model to 57 developmental items.

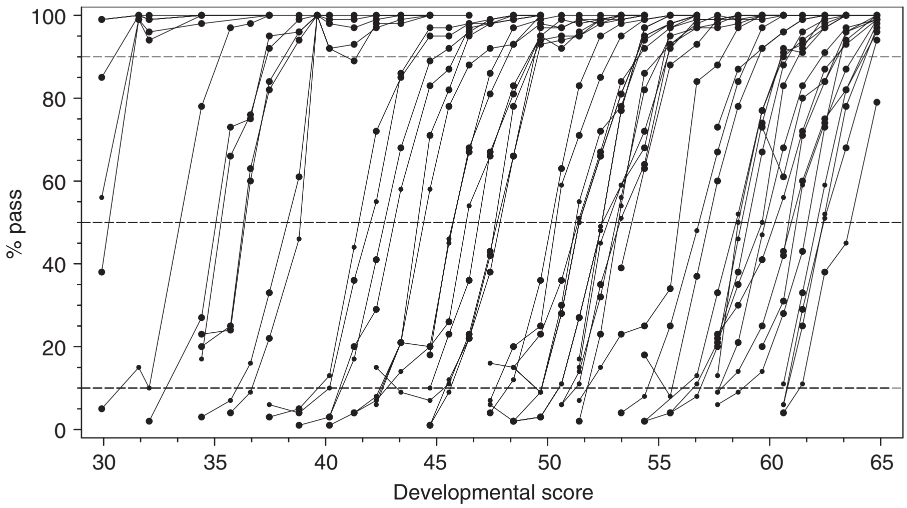


Figure 4. Empirical percentage of passing each of the 57 items of the Van Wiechen scheme by *D*-score.

residual) and under- (high residual) discriminate between persons with a different *D*-score. Some indicators from set 1 have relatively large outfit statistics, but less extreme RUMM statistics. This can be explained by the lower sample size for indicator set 1 (only evaluated during session A), which causes the standardized RUMM statistics to be less extreme. The majority of indicators shows a good fit to the RM. This means that we have a good cause to use person location estimates $\hat{\theta}$ that the analysis revealed as interval scale estimates of development. The percentage of pass scores by *D*-score is shown in Figure 4.

D-scores have almost perfectly parallel associations with percentage pass scores for all indicators, as evident from the curves in Figure 4. A direct consequence is that the relative

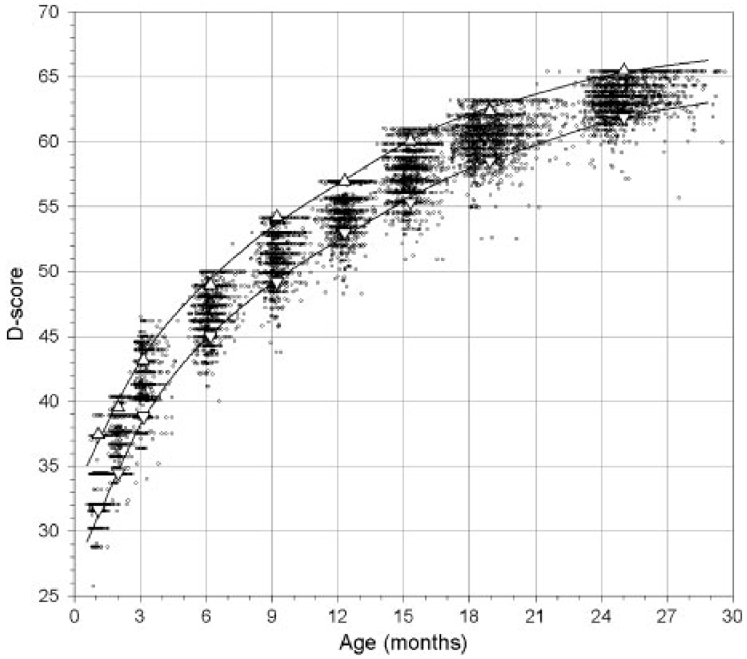


Figure 5. Distribution of the *D*-score by age, including the 10th and 90th centile lines as estimated by the LMS method (both sexes combined).

difficulty of each pair of indicators is the same at different levels of the *D*-score, a prerequisite for the assumption of local stochastic independence. The same step of the *D*-score results in the same increment of the natural logarithm of the probability to pass irrespective of the difficulty of the item.

Figure 5 plots *D*-scores of all children during all sessions against age. The figure also gives age-specific 10 and 90 percentiles of *D*-scores. They were estimated using the LMS method [13]. Worm plots [14] show that the distribution of *D*-scores for a given age is not always very smooth, which is caused by the inherent discreteness of the data. Apart from this fact, estimates appear to be satisfactory. It can be verified from Figure 5 that the LMS-based percentiles (lines) closely matched the empirical *D*-score percentiles (triangles) that we computed within each of the nine sessions.

The new developmental scale brings some interesting new possibilities within reach. The development of children between different sessions can now be expressed as difference scores between successive occasions. This makes it possible to investigate the variability in developmental speed at different ages. This variability can be measured by Pearson correlations of *D*-scores between sessions. See Table II.

The longitudinal nature of the design introduced dependence in the observations. In order to test the effect of it, we randomly selected one session of each child and ran a new analysis. With one session for each child, the indicator-overlap that was shown in Table I still enables estimation of the RM for all indicators and all children simultaneously. The sample size is

Table II. Correlations of developmental scores (*D*-scores) between occasions (time in months).

	B (2)	C (3)	D (6)	E (9)	F (12)	G (15)	H (18)	I (24)
A (1)	0.22	0.20	0.16	0.12	0.14	0.06	0.07	0.11
B (2)		0.30	0.17	0.11	0.14	0.16	0.15	0.11
C (3)			0.29	0.23	0.23	0.17	0.14	0.16
D (6)				0.45	0.36	0.29	0.23	0.22
E (9)					0.50	0.39	0.33	0.29
F (12)						0.59	0.45	0.38
G (15)							0.49	0.40
H (18)								0.49

nine times smaller in this new analysis, because eight sessions of each child were excluded. Results indicate that violation of the independence assumption has not altered the outcomes in our original analysis. The separation index for the new analysis is 0.990, excellent again. The Pearson correlation between *D*-scores (for the subset of cases that was included in both analyses) is 0.9998 and the correlation between indicator difficulties is 0.9990. We did find two differences between the original analysis and the new analysis, which we had anticipated. First, the standard errors of the indicator difficulty estimates are approximately three times bigger in the new analysis. These new standard errors are more realistic, because they are based on one record of each child. Standard errors are relevant for estimating confidence intervals for *D*-scores on the individual level. Second, the RUMM fit residuals range from -2.91 to 2.35 in the new analyses. As a result of the smaller sample size, all RUMM residuals are within the interval between -3 and 3. The relative size of the residuals did not change much. The correlation with original residuals was 0.81.

4. DISCUSSION

We have shown that the measurement properties of the Van Wiechen scheme can be considerably improved. Application of the RM to a longitudinal data set makes a major shift from qualitative observations to quantitative measurements. The reliability measure and fit statistics indicate that the new quantitative *D*-scores succeed in representing outcomes of the Van Wiechen scheme on a common scale. The new scores facilitate better comparability between children as well as longitudinal exploration of development.

To our knowledge this is the first developmental scale for children with such properties. Until now, there was no way to quantify the association between age and development. Exploration of indicator scores by age (Figure 2) suggested that young children up to approximately five months have the fastest development [3]. The underlying assumptions of a common development scale and parallel associations between *D*-score and indicator outcomes were never tested however. The RM aids in testing these assumptions. Application of the RM revealed a *D*-score that has a relation with age (Figure 5) that is quite similar to age-conditional references for height or weight.

The correlations of *D*-scores between occasions in Table II indicate the variability in developmental speed between children [15]. A higher correlation indicates a lower variability of the developmental speed. If, for example, developmental speed were equal for all children, then all children would remain on the same centile within the distribution of children, and the correlation would be 1. Table II reveals that prolonging the time interval between two sessions results in a lower correlation. The correlation between sessions A and I, for example, is lower than the correlation between sessions A and B. So, variability of developmental speed increases with the size of the time interval. Apart from that, the correlations between sessions are higher when children are older. The correlation between H and I, for example, is higher than the correlation between A and B, even though there is only one month between A and B, but 6 months between H and I. These findings precisely match what was shown in previous work by Cole [16] on the development of weight during early childhood. The developmental speed of older children is less variable. The clinical implication of this is that a delayed development at an older age is more likely to persist.

The range of the *D*-score is restricted by the difficulty of the set of indicators that is measured. In our design, this difficulty depends on age, adjusted to the development that children of a certain age normally display. However, we found that the association with age is not just a consequence of our design. In each of the age groups, there are some children (well below 10 per cent) that have positive scores on all indicators, including indicators from the more difficult set that is targeted at older children. Some children therefore hit the ceiling of the scale at the age-specific maximum score, as can be seen in Figure 5. However, minimum scores are not restricted like that; in only 4 cases (0.02 per cent) a child failed all of the indicators. The much lower chance of failing all indicators instead of passing all indicators, is of course a consequence of the 90 per cent success rate that was built in for the age-appropriate indicator set. It means that the least developed children, who are of most interest from a clinical point of view, can be measured with maximal precision.

The introduction mentioned three problems that are associated with current developmental scores. Application of the RM can address each of these. First, the new *D*-scores are not relative to any population or norm group, but only depend on the set of indicators that define the underlying developmental scale. As a consequence, the estimated score of every single measured person may be compared to any other person's score, which deals with our second concern. Third, a child's progress in time can be measured by means of difference in scores between sessions. The *D*-scores represent a substantial improvement for the practice of monitoring development, and create possibilities for more comprehensive analysis of developmental data.

REFERENCES

1. Black MM, Matula K. *Essentials of Bayley Scales of Infant Development—II Assessment*. Wiley: New York, 2000.
2. Van Wiechen W. *Ontwikkelingsonderzoek op het consultatiebureau.* Nationale kruisvereniging, 1988.
3. Schlesinger-Was EA. *Ontwikkelingsonderzoek van zuigelingen en kleuters op het consultatiebureau.* Leiden University, 1981.
4. Den Ouden L, Rijken M, Brand R, Verloove-Vanhorick SP, Ruys JH. Is it correct to correct? Developmental milestones in 555 'normal' preterm infants compared with term infants. *The Journal of Pediatrics* 1991; **118**(3):399–404.
5. Herngreen WP, Reerink JD, Noord-Zaadstra BM van, Verloove-Vanhorick SP, Ruys JH. SMOCC: design of a representative cohort-study of live-born infants in The Netherlands. *European Journal of Public Health* 1992; **2**:117–122.
6. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedogogische Institut: Copenhagen, 1960.

7. Andrich D. *Rasch Models for Measurement*. Sage Publications: Newbury Park, 1988.
8. Van Buuren S, Hopman-Rock M. Revision of the ICIDH severity of disabilities scale by data linking and item response theory. *Statistics in Medicine* 2001; **20**(7):1061–1076.
9. RUMM Laboratories, www.rummlab.com.au, 2003.
10. Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement* 2003; **4**(3):205–221.
11. Wright BD, Masters GN. *Rating Scale Analysis*. Mesa Press: Chicago, 1982.
12. Linacre JM. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 2002; **16**(2):878.
13. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalised likelihood. *Statistics in Medicine* 1992; **11**:1305–1319.
14. Van Buuren S, Fredriks AM. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine* 2001; **20**:1259–1277. DOI:10.1002/sim.746
15. Cole TJ. Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine* 1994; **13**: 2477–2492.
16. Cole TJ. Conditional reference charts to assess weight gain in British infants. *Archives of Disease in Childhood* 1995; **73**:8–16.