1.8 Precision

Stef van Buuren[1,2]
Iris Eekhout[1]
with Manon Grevinga
[1]Netherlands Organisation for Applied Scientific Research TNO, Leiden, 2316 ZL, The Netherlands
[2]University of Utrecht, Utrecht, 3584 CH, The Netherlands

This section shows the properties of the D-score when calculated from short tests. The study of quick tests is useful because it reveals the behaviour of the D-score when the measurement is inherently imprecise.

This section covers:

- Structure of milestone subsets (1.8.1)
- Impact of short tests on D-score (1.8.2)
- Impact of short tests on predicting IQ (1.8.3)

1.8.1 SMOCC DESIGN: STANDARD AND ADDITIONAL MILESTONES

At each visit, the SMOCC study collected scores on a set of *standard milestones* (that about 90 per cent of the children will pass) and a set of *additional milestones* (that about 50 per cent of the children will pass). See Section 1.4.1.2.

The SMOCC dataset covers nine different *waves*. The set of milestones used in the DDI varies per visit. The number of standard milestones varies between 2

TABLE 1.8.1
Number of items administered per wave in the SMOCC data.

Age	Standard	Additional
1m	5	2
2m	2	5
3m	5	6
6m	6	7
9m	7	6
12m	6	6
15m	6	6
18m	6	7
24m	7	7

DOI: 10.1201/9781003216315-8

and 7 on various occasions. The additional milestones equal the standard ones from the next wave.

Table 1.8.1 summarizes the scheduled age for each wave, the number of standard milestones and the number of additional milestones.

Figure 1.8.1 shows the subsets of DDI items administered at each age. For example, at the 1-month visit, the five standard milestones are `ddicmm029` – `ddigm056`, and the two additional ones are `ddicmm030` and `ddifmd002`. At the 2-month visit, the standard milestones are `ddicmm030` and `ddifmd002`, and the five additional ones are `ddicmm031` – `ddigmd057`. And so on.

1.8.2 D-SCORE FROM SHORT TESTS

1.8.2.1 Milestone sets

In the analyses done thus far, we have calculated D-scores from responses on the combined (standard plus additional) milestones. Thus, at the 2-month visit, the D-score was calculated from 2 (standard) + 5 (additional) = 7 milestones.

In daily practice, the set of additional milestones is often lacking. This section explores the impact of using the (smaller) subset of standard milestones on measurement error and prediction.

This section reports and compares three D-scores:

1. D-score from standard milestones;
2. D-score from additional milestones.
3. D-score from all available milestones;

Estimation of 1 is more complicated than for 2 and 3, for the following reasons:

- There are fewer milestones, so the estimate is less precise and more influenced by choice of the prior distribution;
- The standard set contains only easy milestones, which are uninformative for the majority of children.

1.8.2.2 Milestone sets at month 2

The vertical axis of Figure 1.8.2 shows the D-score, separately calculated from the standard, additional and all milestones for children aged two months. The colour of the dots represents the number of FAIL ratings within each set of milestones.

At month two there are just two standard milestones: `ddicmm030` and `ddifmd002`. About 90 per cent of the infants will pass these. The green dots in the left-hand panel represent the estimated D-scores corresponding to two passes. As explained in Section 1.5.3.2, we calculate the D-score with an age-dependent prior. If the ages vary (and they do), then the D-score for infants having the same total score will also vary.

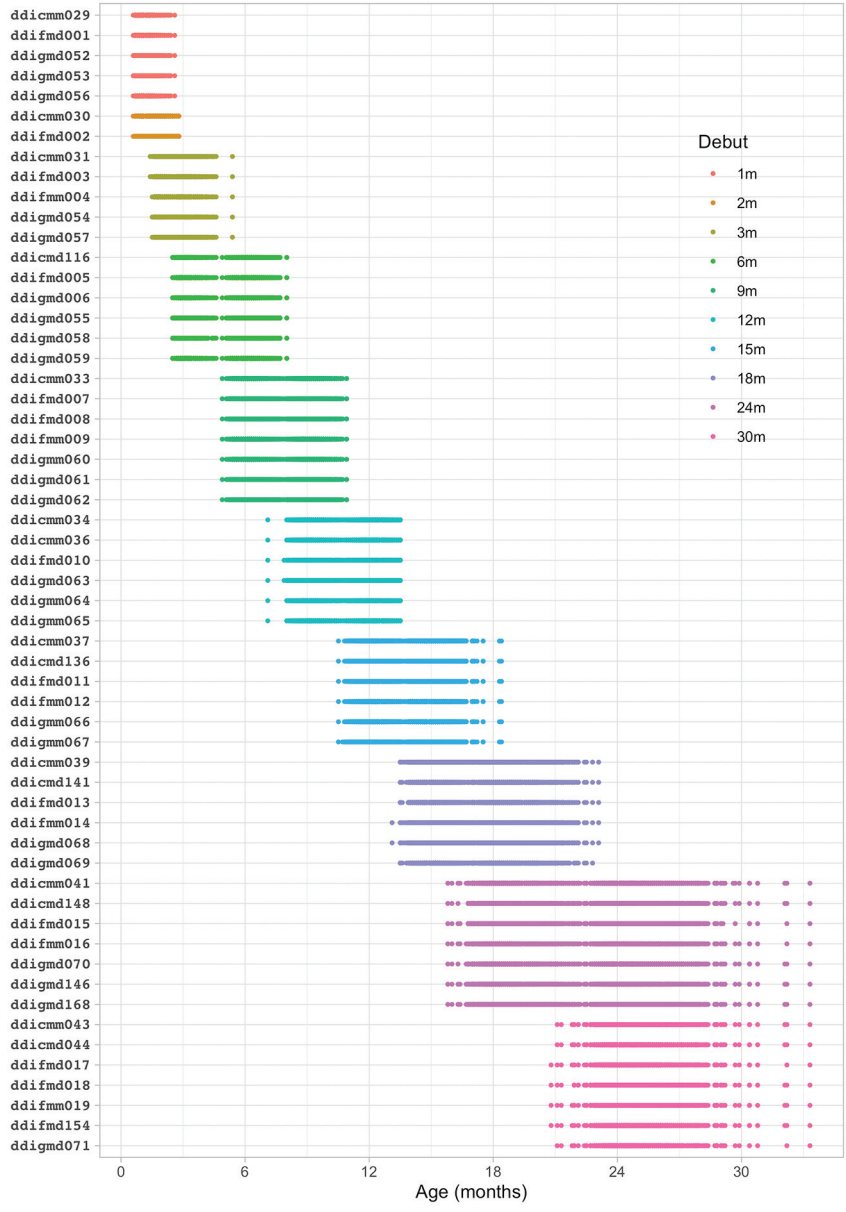

FIGURE 1.8.1 Age-item grid of the SMOCC study, illustrating how the 57 DDI items are distributed over nine visits during the first 24 months.

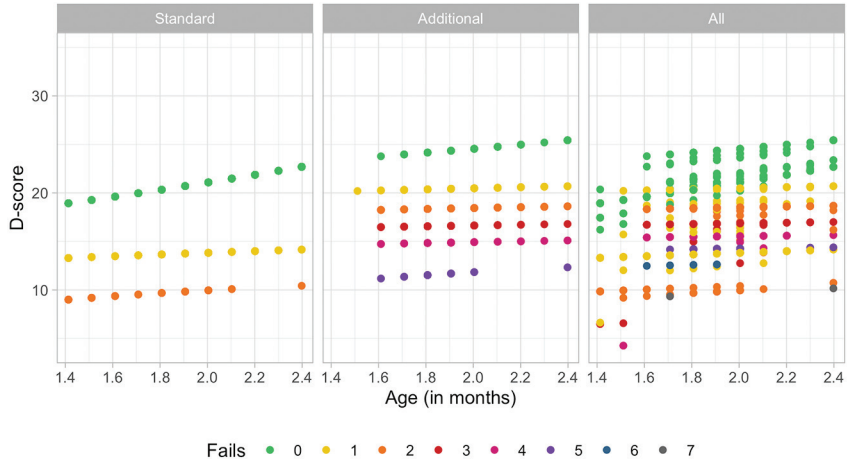

FIGURE 1.8.2 Distribution of the D-scores calculated from the standard, additional and all available milestones at month 2. Colors correspond to the number of fails.

If a child fails either `ddicmm030` or `ddifmd002`, then the D-score is substantially lower. The left-hand figure shows a *gap* between the green dots (perfect score) and the yellow dots (one FAIL). The impact of a FAIL on the D-score is substantial. For example, the D-score of an infant with one FAIL on a standard milestone drops from about 20 *D* to 14 *D*. Thus, with these two milestones, there cannot be a D-score in the range 15 *D* - 18 *D*. It depends on the purposes of the measurement if this is acceptable. We can prevent gaps by measuring more milestones, e.g., milestones taken from the additional set. Another gap occurs between 14 *D* and 11 *D*. These gaps illustrate that precision is constrained if we administer only two milestones.

The middle panel shows the estimated D-score at the same visit but now calculated from the five additional milestones (i.e., the standard milestones from month 3). Infant aged two months have approximately a 50 per cent chance of passing each. Note that administration of the additional milestones will cover the range 14D-20 *D* quite well. Note the ceiling is also higher with these milestones.

Note that the range of the estimated D-scores is quite similar in both plots. This similarity is a result of accounting for the difficulty level of milestones. The estimate of the D-score is *unbiased* for difficulty.

The panel on the right-hand side provides the D-score calculated from all milestones. We can easily recognize the points coming from the standard and additional sets. Also, there is a limited number of ratings on easier items that belong to month 1. We rescored these because the child failed these milestones at the previous visit. Rescoring effectively extends the range of possible D-scores to the lower end, so now we can find some children who have D-score lower than 10 *D*.

1.8.2.3 Milestone sets at month 3

Figure 1.8.3 is the same plot as before, but now for month 3. Compared to Figure 1.8.2, all points shifted upwards because the children are now one month older.

The additional milestones from month 2 are the standard milestones of month 3. In Figure 1.8.2, there were at least 11 children (in purple) failed all five additional milestones. One month later, one child has five fails.

1.8.2.4 Floor and ceiling effects

Figure 1.8.4 plot the D-score distribution for all occasions. Some observations:

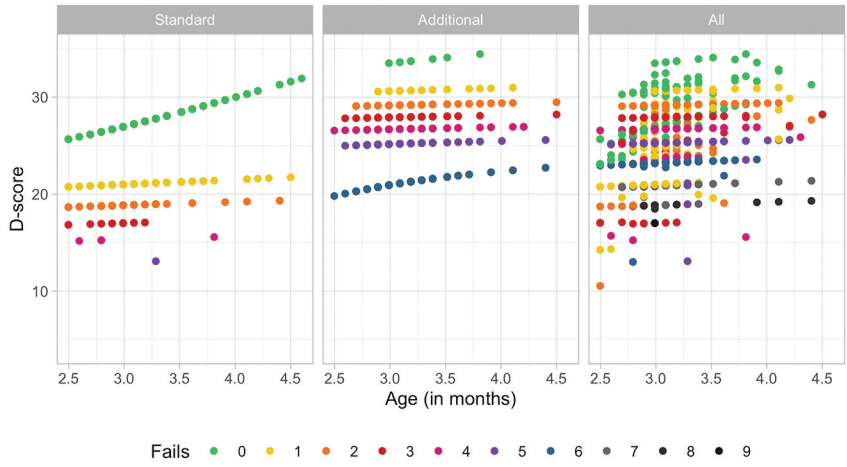

FIGURE 1.8.3 Distribution of the D-scores calculated from the standard, additional and all available milestones at month 3. Colors correspond to the number of fails.

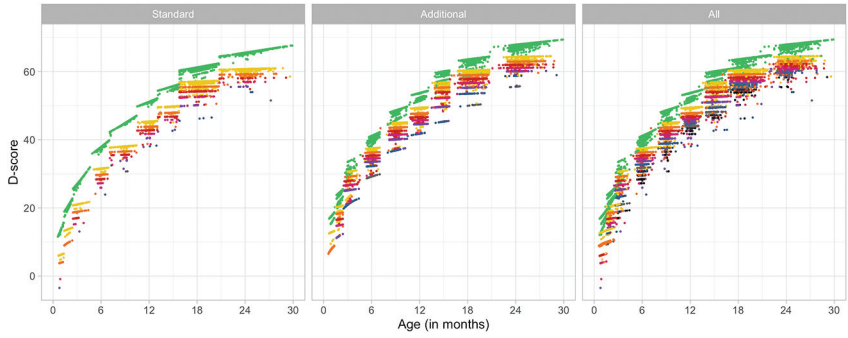

FIGURE 1.8.4 D-score by age 0–30 months for standard, additional and all available milestones at each measurement occasion.

- *Ceiling effect*: The ceiling effect (green) is most prominent in the *standard* set, but is also present in the other two sets. None of the three sets can filter out children with really advanced development. To achieve more precision at the upper end, we would need to include more difficult milestones.
- *Floor effect*: There are almost no floor effects in the *standard* and *all* sets. These sets discriminate well among children with delayed development, which was the designed purpose of the DDI. Note that floor effects are visible in the *additional* set.
- *Average level*: All three sets capture the overall relation between age and development. The *additional* set is quite efficient for measuring average levels development but lacks detail on the extremes.

Figure 1.8.4 shows that a short test (5–6 milestones) can precisely measure the lower tail of the D-score distribution (*standard* set) or the middle of the D-score distribution (*additional* set), but cannot do both at the same time.

1.8.3 IMPACT OF SHORT TESTS ON PREDICTING IQ

1.8.3.1 Measurement and prediction

In Section 1.8.2, we saw that a short test can measure the middle or one tail of the distribution, but cannot be precise for both at the same time. If we want to identify children at risk for delayed development, we are interested in the lower tail of the distribution, so in that case, the *standard* set is suitable. But what set should we use if we want to predict a later outcome?

This section explores that effect of taking different milestone sets on the quality of prediction.

1.8.3.2 UKKI

Hafkamp-de Groen *et al.* (2009) studied the effect of the D-score on later intelligence, using a subset of 557 SMOCC children that were followed up at the age of five years.

The Utrechtse Korte Kleuter Intelligentietest (UKKI) (Baarda, 1978) is a short test to measure intelligence. The UKKI is a simple test with just three components:

- Redraw five figures (square, triangle, cross, trapezoid, rhomboid);
- Draw human figure, with 28 characteristics, like legs, eyes, and so on;
- Give meaning to 13 words like knife, banana, umbrella, and so on.

Administration time is about 15–20 minutes. The UKKI has a reasonable test-retest reliability for group use (Pearson $r = 0.74$, 3-month interval).

1.8.3.3 Exploratory analysis

Figure 1.8.5 shows the empirical IQ distribution of 557 children. The mean IQ score is 108, and the standard deviation is 15, so the IQ scores of children in the sample is about a half standard deviation above the 1978 reference sample.

Figure 1.8.6 shows that the relation between the D-score 0–2 years and IQ at five years is positive for all milestone sets and all ages. The strength of the association increases with age. At the age of 2 years, the regression coefficient for D-score is equal to β (D) = 1.4 (SE: 0.21, $p < 0.0001$), so on average an

TABLE 1.8.2
Pearson correlation between D-score (0–2 years) and IQ at 5 years.

Visit	Standard set	Additional set	All milestones
1m	0.059	0.005	0.027
2m	0.051	0.056	0.048
3m	0.036	0.100	0.102
6m	0.040	0.038	0.036
9m	0.094	0.143	0.132
12m	0.046	0.162	0.137
15m	0.180	0.153	0.187
18m	0.129	0.153	0.146
24m	0.245	0.255	0.267

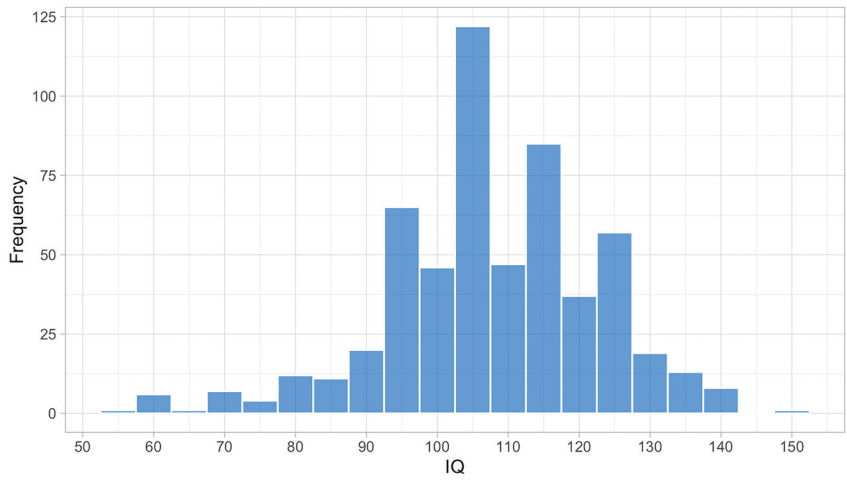

FIGURE 1.8.5 Histogram of UKKI *IQ* scores taken around the age of five years (SMOCC data, n = 557).

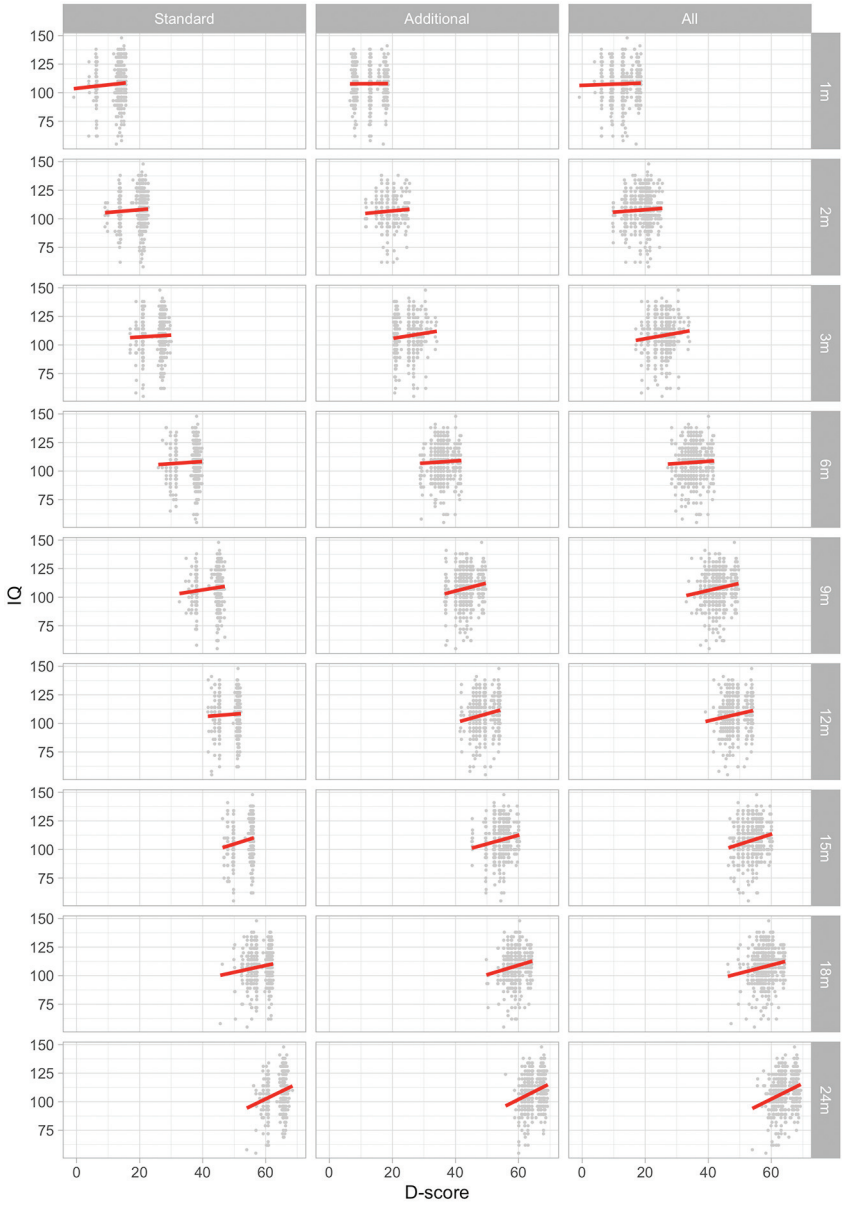

FIGURE 1.8.6 Relation between D-score at infancy and *IQ* at age 5 years according to three milestone sets and nine visits (SMOCC data, $n = 557$).

increase of 1.0 unit in the D-score at the age of 2 years corresponds to a 1.4 IQ-score points increase at the age five years.

Table 1.8.2 summarizes the Pearson correlations between the D-score and later IQ. The association between D-score and IQ is weak during the first year of life but gets stronger during the second year. In general, having more (and more informative) milestones helps to increase the correlation, but the effects are relatively small. So even from the standard set of the seven easy milestones at 24m, we obtain a reasonable correlation of 0.245.

All in all, these results suggest that neither the amount nor the difficulty level of the milestones is critical in determining the strength of the relation between the D-score and IQ.