APPENDIX 1

Analysis of Variation in Inter-Measurement Agreement

Variation in inter-measurement agreement as a function of measurement amplitude was analyzed according to the regression method of Bland and Altman (manuscript reference 27). First, linear regression was performed for the difference between measures against their average, describing systematic bias between first and second measurements as a function of measurement amplitude. The result, henceforth termed the "bias function," is the regression fit to a standard Bland-Altman plot (e.g. middle dotted line in Figure 5 of manuscript). If the slope of this function is significantly different from zero, bias varies with measurement amplitude. Otherwise, the mean bias can be described as a constant, the average of all inter-measurement differences, and a t-test can be used to assess its significance.

The inter-measurement differences predicted by the bias function are then subtracted from those actually observed. The resulting residuals take into account any systematic differences between the measurements being compared, and thus reflect inter-measurement variability alone. Linear regression is performed for the absolute values of the residuals against measurement amplitude. The resulting function, herein termed the "variability function," describes inter-measurement variability as a function of measurement amplitude. If the y-dimension scatter on a Bland-Altman plot increases along the x-axis, this is reflected as a positive slope of the variability function. If the slope is significantly different from zero, repeatability varies with measurement amplitude.

For any given measurement amplitude, the mean inter-measurement difference is predicted by the bias function. The SD around this mean is predicted by the variability function, multiplied by a factor of $\sqrt{(\pi/2)}$. For any given measurement amplitude, an amplitude-specific RI is defined according to the definition of RI: 1.96 times the SD of the inter-measurement difference.

Graphically, the 95% limits of agreement are represented by linear functions describing (mean difference +1.96 SD) and (mean difference - 1.96 SD) as a function of measurement amplitude (e.g., sloping dotted lines in Figure 5 of manuscript).

The same methodology is applicable to the study of inter-measurement variability as a function not only of measurement amplitude, but of any parameter of interest, and thus was also used to assess whether repeatability varied significantly with age or refractive error. The bias functions described above were also assessed for inter-measurement differences as a function of measurement amplitude when comparing lag measurements using different dynamic retinoscopy methods.

The Table in Supplemental Digital Content 2 summarizes the findings of these analyses, and the tests of whether the slopes of bias and variability functions are significantly different from zero.

Impact of Testing Sequence; Verification of Linearity of Accommodative Demand-Response Functions

The testing sequence for different retinoscope distances (33 cm, 50 cm, 67 cm, 33 cm, 50 cm, 67 cm) might impact accommodative demand-response slopes through fatigue or practice effects. The mean difference between successive lag estimates at a given retinoscope distance was assessed in the screening cohort. It was 0.15 D (SD 0.59 D), 0.04 D (SD 0.41 D), and -0.02 D (SD 0.32 D), for 33 cm, 50 cm and 67 cm retinoscope distances, respectively. The 33 cm position showed a small but significant decrease in lag with repeated testing ($t(163)=3.26$; $p=0.001$), but other positions did not ($p>0.20$).

To evaluate the linearity of accommodative demand-response functions, an average "low-response" slope was calculated for each child in the screening cohort, using the data from 67 cm and 50 cm retinoscope distances, and a "high-response" slope using data from 50 cm and 33 cm retinoscope distances. The mean difference between individual children's low-response and high-response slopes was 0.04 (s.d. 1.11), not significantly different from zero ($t(167)=0.47$; $p=0.64$), indicating no systematic non-linearity in accommodative demand-response functions.

Repeatability Analysis: Pooling of Screening and Clinic Cohorts

First and second MBR standardized 40 cm target estimates the same day were correlated, for both screening and clinic (Day 1) cohorts: Pearson's $R=0.79$ ($t(170)=16.7$, $p<0.0001$) and $R=0.86$ ($t(26)=8.7$, $p<0.0001$), respectively. Two outliers whose cooperation level changed between measurements were excluded from the clinic cohort. Measured with 4 D and 2.4 D uncorrected hyperopia, respectively, their initial Day 1 lag estimates were 3.8 D and 4.4 D, while subsequent Day 1 estimates and all Day 2 estimates were between 1.1 and 1.7 D.

Stratified according to the amount of lag, the screening and clinic cohorts' repeatability indices (RI) were comparable: 0.33 D (95% CI 0.30-0.38 D) and 0.29 D (95% CI 0.20-0.45 D), respectively, for lag <0.5 D, and 0.61 D (95% CI 0.51-0.77 D) and 0.90 D (95% CI 0.64-1.42 D), respectively, for lag ≥0.5 D. The cohorts were therefore pooled for further analysis.

Comparison of Orthogonal and Linear Regression for Standardized 40cm Target Estimates

Orthogonal regression was used for the primary analysis of accommodative demand-response functions, to derive standardized 40 cm target estimates, but ordinary linear regression of accommodative response (y) on accommodative demand (x) was also performed for comparison. In the screening cohort (first estimate), the mean difference between lag estimates

using linear regression and orthogonal regression was 0.004 D (SD 0.016 D). The methods agreed within +/- 0.03 D. For clinic cohort measurements (Day 1 or Day 2, first or second estimate), the mean difference between estimates using linear regression and orthogonal regression (excluding one outlier) ranged from 0.01 D to 0.03 D (SD range 0.06 D to 0.12); the methods agreed within +/- 0.12 D to +/- 0.23 D.

The within-visit inter-measurement RI for screening and clinic cohorts pooled, using linear instead of orthogonal regression, was 0.35 D (95% CI 0.31-0.39 D) for lag <0.5 D, and 0.65 (95% CI 0.55-0.80) for lag ≥0.5 D, similar to results using orthogonal regression.