The desire to create and validate an instrument that can be used with confidence to conclusively demonstrate the effectiveness of treatment in infants receiving physical therapy is a laudable goal. Palisano et al have shown the merit of a sequential, step-by-step process in establishing the measurement properties of the Peabody Developmental Gross Motor Scale (PDMS-GM) for this purpose. They describe a detailed study in which they answer key questions for evaluative measures: Does the measure show changes in group scores over time? Do most infants show changes in their individual scores? Is the measure responsive to clinically important differences in gross motor skills, and can data from the study be used for sample-size calculations in clinical trials? This article, however, should be making only tentative conclusions for a variety of methodological reasons.

In establishing validity of an evaluative measure, it is commonly agreed that a new measure must demonstrate its ability to record changes in subjects who are actually changing. Without a "gold standard" to assess whether individuals in the sample are truly changing, one cannot conclude that the change scores that may be observed actually reflect true change.1 One may merely be observing measurement error or random variation in gross motor skills unless the study also collects data from an external measure that is applied concurrently with the test measure. Alternatively, one could divide the sample into groups known to be "stable" and "changing" by some other criterion, such as parent or therapist report, and then compare the ability of the test to show differential change between the groups.2,3 Finally, one could apply the test measure to a sample that receives an intervention of known efficacy and observe whether the test measure correctly detects a change in motor skills.

The authors have not incorporated any of these comparative techniques in the study, and thus we are left wondering whether the observed change scores represented true change in motor skills. They appear to rely on the fact that the sample was recruited from children receiving therapy, as if that guaranteed that the children would actually be changing their skills. Even this assumption, however, does not appear to hold for the reported 38% of children who did not record a criterion reliable change index value of 1.96 and thus whose change scores may have been due to measurement error. It is questionable whether the authors' conclusion that the PDMS-GM can detect true change in motor skills in the diagnostic groups can be fully supported.

A similar lack of comparative information is evident in the sample selection. The authors use the PDMS-GM itself to classify children with motor delay. They also report, however, that the measure has not been validated for this purpose in the 0- to 5-month-old population. If the PDMS-GM is not sensitive in discriminating delay in this
age group, then the sample may have included children without motor delay; thus increasing the likelihood of reporting a change in PDMS-GM scores for the subgroup. Perhaps the authors could have used a different measure, such as the Bayley Motor Scales, to establish motor delay in this sample.

Guidelines for establishing reliability in evaluative measures also suggest that test-retest reliability is of paramount concern or else observed change may be due only to intrasubject variability. Apparently, test-retest reliability was conducted in this study with only 9 of 124 subjects, which leaves doubts about the completeness of the reliability testing.

The principal methodological concern in this study relates to the issue of responsiveness, or the power of the PDMS-GM to detect a clinically important change in motor skills when change has actually occurred. A responsive measure should show little variability due to time in a "stable" group and greater variability due to time in a "changing" group. The practical issue here is in being able to estimate the sample size required in a trial to demonstrate an expected effect size in a study population.

The problem in the article is in two areas: (1) estimating the size of Delta, or the minimal clinically important change, and (2) selection of the sample that was used to calculate the responsiveness index. The authors used a small sample of five infants with unknown characteristics, observed by two therapists, to derive an estimate of Delta = 10, or a 1-month gain in age-equivalent score on the PDMS-GM. The study findings, however, indicate that the entire sample's mean change in age-equivalent scores varied from 2.2 to 4.0 months. Assuming that this range approximates a minimal clinically important change, the authors appear to have underestimated Delta and thus overestimated the sample size required in a clinical trial using the PDMS-GM.

The more important issue, however, is in the use of the entire sample of children receiving treatment to calculate the denominator of the responsiveness index. According to the originators of the method that the authors use, the most appropriate measure of responsiveness would relate the clinically important difference achieved by subjects who are changing (Delta) to the variability in test scores in stable subjects. Thus, Delta should be calculated from data obtained from those who are actually changing, and the authors have tried to do this. However, the denominator, variability in test scores (the square root of twice the mean square error), should be calculated from data obtained from only those who are stable. By including all subjects, both stable and changing, in the denominator, the mean square error is likely inflated and the sample size is overestimated again. A solution would be to sort subjects into "stable" and "changing" groups and apply an analysis of variance approach using repeated measures on the initial and follow-up scores of the "stable" group. These methodological problems call into doubt the conclusion that the PDMS-GM can only be used in larger studies of treatment effectiveness. Considering the difficulty in recruiting suitable samples of children for clinical trials, the authors' conclusions may needlessly preclude the use of the PDMS-GM in such research.

These methodological points illustrate the problem with utilizing single measures in validity studies. By their nature, validity studies imply a comparison of two measures or sample groups to determine whether the test instrument actually measures the construct of interest. Without using methods to properly select comparison groups or judge the test's ability to achieve its stated purpose, one is comparing the measure with itself.

The authors acknowledge that the involvement of clients and their families in this process of test validation is crucial. In pediatric research, parents can play a major role in estimating a minimal clinically important change from a different and comparative perspective. There are of course problems in implementing such judgments of change and not all efforts are successful, depending on the complexity of the construct being assessed.

Overall, the authors are conservative in their assessment of the suitability of the PDMS-GM as an evaluative tool for infants receiving therapy. Although I support the conclusion that more specific measures are needed to evaluate the efficacy of treatment in specific populations, the methodological problems in this study raise doubts concerning even these conservative conclusions about the PDMS-GM.

William F Boyce, MSc, PT
Assistant Professor
School of Rehabilitation Therapy and the Departments of Paediatrics and Community Health and Epidemiology
Queen's University
Kingston, Ontario, Canada K7L 3N6
boycew@post.queensu.ca

References
Invited Commentary
William F Boyce
PHYS THER. 1995; 75:948-949.