

## Ten Top Problems with Normed Achievement Tests for Young Children

Dumont, R. & Willis, J., Fairleigh Dickinson University

10. There is no agreed-upon preschool or kindergarten curriculum at national, state, and even, in some cases, local levels. It is difficult to sample a curriculum that does not exist. For higher grades, there is at least some commonality among the many curricula at a given grade level. The same skill may be placed at very different levels. See, for example, [http://alpha.fdu.edu/psychology/WR\\_vs\\_LWI.htm](http://alpha.fdu.edu/psychology/WR_vs_LWI.htm), [http://alpha.fdu.edu/psychology/NO\\_VS\\_CALC.htm](http://alpha.fdu.edu/psychology/NO_VS_CALC.htm), and [http://alpha.fdu.edu/psychology/MR\\_VS\\_AP.htm](http://alpha.fdu.edu/psychology/MR_VS_AP.htm)
9. Young children are often inconsistent in their responses, which would argue for more items to increase reliability.
8. But young children have short attention spans and they fatigue easily, which requires fewer items.
7. Sampling works well for a large domain. For example, if a child is expected to have a reading vocabulary of 3,000 words, it is pretty easy to estimate reading skill with a 25-word test. However, if a child is expected to have a reading vocabulary of 10 words, your 25-word test could, by pure chance, easily sample all 10 or none of them, giving an inflated or depressed estimate of the child's reading ability. Similarly, many achievement tests for young children have only a few letter-naming items, rather than 52. If a child knows ten to sixteen letters, a ten-item test could easily hit or miss all of them by pure chance. If you test a child on Monday, and the teacher teaches the vowels on Tuesday, that could be the difference between a score of zero and a score of five on the ten-item test.
6. Young children develop new skills so rapidly that norms tables should be divided by weeks, not three, four, or six months. The difference between age 10-0-0 to 10-6-29 and age 10-7-0 to 10-11-29 may be trivial, but the difference between 4-0-0 to 4-6-29 and 4-7-0 to 4-11-29 is tremendous.
5. Item format matters a lot more for younger children. Most ten-year-olds don't care whether an addition problem is presented horizontally or vertically, but five-year-olds do. The space between lines on writing paper and the presence or absence of a dotted midline is a deal-breaker for most kindergarten students.
4. Item gradients are necessarily very steep for younger children. There aren't any clearly defined steps between not being able to write the letter M and being able to write it.
3. Norming samples are also a huge problem at the preschool level. If you carefully sample geographic regions, parental education and income, and other germane variables, you can be fairly safe in using whatever public and private schools (in the right proportions) are available. However, at the preschool level, there is a huge difference between the Mary Poppins School of Unfettered Self-Expression and Free Play and the John Stuart Mill Preschool of Relentless Academic Pressure. Low-income kids from the JSMPRAP are likely to score higher on academic achievement tests than rich kids from the MPSUSEFP. A truly representative national sample (especially with only 100 to 300 kids per year of age) is virtually unobtainable.

2. There often is insufficient floor for young children on achievement tests. See, for example, Goldman (1989) and <http://alpha.fdu.edu/psychology/McGee.htm>.
1. Consequently, formal and informal, criterion-based tests (with exhaustive sampling, e.g., all 52 letters, all sums up to ten, etc.); curriculum-based measurement; classroom observations; and work samples are likely to be much more informative than normed tests up to at least a mid-second-grade level of achievement (regardless of actual grade placement).

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*, (2), 155-166.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484-502). New York: Guilford Press.

Goldman, J. J. (1989). On the robustness of psychological test instrumentation: Psychological evaluation of the dead. In Glenn G. Ellenbogen (Ed.) *The Primal Whimper: More Readings from the Journal of Polymorphous Perversity*. New York: Ballantine, Stonesong Press.

Schultz, M. K. (1988). A comparison of Standard Scores for commonly used tests of early reading. *NASP Communiqué, 17* (4), 13.