Review of **Metropolitan Integrated Performance Tasks**, by Theodore Coladarci[1]

The Metropolitan Integrated Performance Tasks (MIPT) are designed to engage students in a series of activities that "show how each child is progressing in acquiring the concepts, strategies, and skills needed to perform language/literacy and quantitative tasks typically introduced in PreKindergarten and Kindergarten" (scoring guide, p. 5). Unlike conventional tests, these performance tasks are intended to "mirror active, hands-on instruction." For example, a component of one task asks students to match a numeral with the correct number of objects. This is accomplished by having students (a) cut out 15 large dots from a sheet of paper; (b) examine a page in the test booklet that presents a row of numerals, each with an empty box above it; and (c) paste the correct number of dots in each box. A component of another task involves the sequencing of story events: Students cut out pictures of events from a story that had been read to them moments before, and then they paste the pictures in the test booklet in proper chronology. The teacher plays a facilitative role, in that he or she "should motivate, guide, and encourage children to produce their best work."

There are two levels of the MIPT; each level offers two tasks, and each task comprises six activities. Level 1 tasks correspond to objectives usually taught in prekindergarten and early kindergarten; Level 2 tasks reflect those objectives taught toward the end of kindergarten or the beginning of first grade. All tasks are "thematic." For example, activities for the Level 2 task Birthday Surprise are situated in a story of a girl who is selecting a birthday gift for her grandfather.

The Metropolitan Integrated Performance Tasks are designed to assess two general sets of objectives: "Language/Literacy" (e.g., follow oral directions, copy print, classify objects, predict a story ending, sequence story events, tell a story from pictures, comprehend a story, identify beginning consonants and consonant blends, understand positional words) and "Quantitative/Mathematics" (e.g., count, understand numerical sequence, demonstrate number/numeral correspondence, understand ordinal numbers, measure objects in a picture, understand relational concepts, compute, understand fractions, complete number sentences, write and solve a story problem, extend patterns).

Performance task activities can be spread over several days, or even weeks. Further the MIPT can be administered to any size group, from one child to the entire class. A few activities have "cooperative" components.

ADMINISTRATION. Extensive directions are provided for administering the performance tasks: Teachers are told what materials are needed for each task, how to introduce the task theme, and how to conduct the task's six activities. My sense is that many teachers will find the administration of the MIPT to be, at best, logistically daunting. First, each activity has multiple steps and instructions; with a classroom of young children, complications and confusion easily arise. Second, teachers are instructed

[1] Coladarci, T. (2001). Review of the Metropolitan Performance Assessment: Integrated Performance Tasks. In J. C. Conoley & J. C. Impara (eds.), *The Fourteenth Mental Measurements Yearbook* (pp. 741-743), Lincoln, NE: The Buros Institute of Mental Measurements.

to "circulate" among students to query each child about his or her work and, if necessary, to write clarifying information in the test booklet regarding a student's response. Third, an accompanying observation form requires teachers to note each student's behavior during the activities (e.g., works without being distracted, tries a variety of solutions, takes turns).

SCORING. Each of the four tasks yields a single "Holistic" score, as well as an "Analytic" score for Language/Literacy and for Quantitative/Mathematics. All scores fall on a 3-point scale. The Holistic score "represents the scorer's overall impression of a child's performance" (scoring guide, p. 7) across the task's six activities. That is, does the student demonstrate "the strategies, concepts, and processes" relevant to the task's activities? Holistic scores take on values of 3 (all or almost all), 2 (some), or 1 (few or none). In contrast, Analytic scores reflect the student's mastery of area-specific objectives. On a Level 1 task, for example, a Quantitative/Mathematics Analytic score of 3 is reserved for students who (a) demonstrate understanding of number/numeral correspondence, (b) sequence numbers and sets correctly, (c) graph toys correctly, and (d) record the total numbers of graphed toys accurately. This score signifies that "performance in this area is successful." An Analytic score of 2 reflects "a combination of strengths and weaknesses" in the area, and a score of 1 corresponds to "largely unsuccessful" performance. Data from the observation component are not considered in determining these scores.

For each task, a scoring guide provides detailed rubrics for both Holistic and Analytic scoring, along with "hints" and "tips." Sample responses, illustrating various levels of proficiency, are provided for each task. Holistic and Analytic scores accompany each sample response, along with explanatory annotations. Finally, teachers can score a practice exercise and compare their scores to those of experienced raters. The MIPT author, Joanne Nurss, is to be commended for assembling this rather impressive Scoring Guide. Nevertheless, I suspect that many teachers will question whether several 3-point scores are worth the tedium of administering the MIPT.

RELIABILITY. Interrater agreement was examined with raters trained at The Psychological Corporation. Exact agreement in Holistic scoring was obtained for 82% to 90% of these raters, depending on the task; the figures for Analytic scoring were 81% to 86%. Agreement was always within 1 point on these 3-point scales. A subsequent analysis was conducted using novice raters (classroom teachers). No further information is provided about this analysis, other than that "agreement [between these teachers and] the trained scorers was nearly as high as that between trained scorers" (emphasis added). I take this to mean that some of the exact agreement percentages may have dropped below 80%, which is rather low when there are only 3 score points. These reliability data suggest that the MIPT scoring rubrics and explanatory annotations should more clearly differentiate adjacent performance levels.

VALIDITY. The validity argument for the MIPT is thin. In regard to content validity, we are told only that the objectives and content of this instrument were informed by an "extensive review of the literature" concerning developmentally appropriate curricula, emerging literacy, and the like. Although Nurss acknowledges that the content of the MIPT should be of demonstrable relevance to "the processes and strategies important for success in beginning reading, writing, numeracy, and problem solving" (scoring guide), this relevance is not demonstrated to the prospective user. As for construct validity, Nurss implies that the face validity of the MIPT is sufficient evidence, "because the tasks and the behaviors they measure are exactly the same." This betrays an unusual definition of construct validity and, in any case, does nothing to assure prospective users that MIPT scores permit meaningful inferences about the knowledge and skills measured by this instrument. For example, a Level 1 activity asks students to match uppercase and lower case letters by connecting a series of labeled dots (A to a, B to b, etc.), which ultimately forms the outline of a familiar object. Does this problem reveal a student's knowledge of uppercase and lowercase letters or, rather, the ability to complete a picture? In some instances, the mere configuration of dots makes a correct response seemingly unavoidable, even for students who do not know their Ps and Qs.

The author's fondness for face validity notwithstanding, correlations between MIPT performance and Metropolitan Readiness Tests (MRT) scores are available--although, surprisingly, one must order the MRT norms book to see them. The Level 1 MIPT holistic score and the Level 1 MRT Total Test Composite are moderately correlated (r = .51, n = 76), as are the Level 2 tests (r = .54, n = 49); correlations between MIPT analytic scores and MRT subscale scores are generally smaller (rs = .29 to .53). These validity coefficients are somewhat modest, although the restricted scale of MIPT scores doubtless is at play.

No evidence of predictive validity is provided because, curiously, "predictive validity is not relevant" (scoring guide) to the MIPT. Although the MIPT was not designed for making formal predictions about subsequent performance, one nonetheless would expect that MIPT scores are related to academic progress. Evidence of this should be furnished.

Nurss embraces the important notion of consequential validity, which she defines as "the extent to which an assessment leads to improvement of instruction and learning in the classroom" (scoring guide). Surprisingly, no evidence is provided in this regard, nor is there even any discussion of how the MIPT can be used to inform instruction. In short, the instructional value and utility of this instrument remains undemonstrated.

CONCLUSION. Reliability and validity evidence for the MIPT is generally weak. Further, although it is true that many of the performance tasks have the feel of authentic activities, this authenticity, perhaps ironically, calls into question the need for such an instrument. Because most early education teachers routinely engage students in such activities, there is little reason to believe that the MIPT will yield information about students that their teachers do not already know. To be sure, some teachers may not be accustomed to deriving general ratings from naturally occurring instructional activities. But rather than purchase off-the-shelf assessments that duplicate instructional practice, a

school district would be better advised to invest in staff development (e.g., constructing and using scoring rubrics) so that teachers can extract reliable and valid information from the classroom activities in which they routinely engage their students.