Quality of Performance and Change-Sensitive Assessment of Cognitive Ability

Gale H. Roid, Ph.D.
Washington State University, Vancouver, USA

Growth and change are fundamental processes in human development. Recent trends in education, psychology, medicine, and other fields have stimulated renewed interest in methods of measuring developmental growth and change (Collins & Sayer, 2001). Educators and governmental agencies have recently placed greater emphasis on the importance of growth in reading as foundational to student learning (Torgesen, 2002). A major review of the regulations is underway in the United States, concerning funding of special education in public schools (e.g., President's Commission on Excellence in Special Education, 2002) and the emphasis has been on "continuous progress" assessment and accountability for student learning (e.g., multiple testing during the school year). Extensive research is underway on the developmental time tables in antisocial behavior (Bauer & Estell, 2001). Highly sophisticated multivariate statistical models are being developed for assessing change in medical programs directed at drug-use prevention, depression recovery, and head-injury recovery (Collins & Sayer, 2001). Measures of growth are also needed in monitoring the progress of infants born prematurely (Krishnakumar & Black, 2001; Roid & Sampers, 2004). When cognitive performance decreases rather than increases, as in the elderly (e.g., memory function), measures sensitive to change in the negative direction are needed (Roid, 2003b).

Definitions

Growth refers to any incremental improvement in cognitive functioning, however small. Growth is most obvious with repeated, individual (longitudinal) testing. Increments of growth are analogous to the changes in performance noted across age groups, from birth to adulthood, as measured by growth curves of test scores. Change, in the context of the current paper, means any increment of improvement, decline or recovery in cognitive functioning. This change may be due to a variety of causes, including typical cognitive development, injury or illness, or response to treatment or intervention. Change-sensitive assessment refers to any evaluation, based on test scores and other information about an individual that is collected or studied at two (or more) points in time and used to evaluate growth or change. Change-sensitive assessments are particularly helpful in evaluating learning capacity, response to intervention, effectiveness and appropriateness of treatment and general tracking of growth or change in an individual across time.

Note. Paper presented as the keynote address for meetings of the International Test Users Conference, Melbourne, Australia, July, 2004; Revised, October, 2004 for the International Test Commission meetings, Williamsburg, Virginia.

Quality of performance methods refer to testing or observational procedures that are designed to identify small increments of difference in the quality (not just quantity or presence/absence) of actions, behavior, performances, or products created by the target individual being assessed. Quality-of-performance measures help in the identification of borderline, or mild, developmental delays because children may obtain "milestones" (behavior occurring at the expected age such as walking by age 1 year) but with unexpected quality or atypical characteristics.

The Role of Item Response Theory in Change-Sensitive Assessment

Major measurement tools for building change-sensitive assessments have been developed through research on item response theory (IRT) models. IRT models (Lord, 1980) are a large family of mathematical models used to analyze test items, develop collections of items, create scales, and produce test scores for examinees. Following decades of research on IRT models, Embretson (1996) recently asserted that the models had reached such an acceptable level of scientific verification that they should replace classical test theory (e.g., Gulliksen, 1950) as the "new rules of measurement" in psychology and education.

The version of item response theory that I have applied most often to individually-administered tests in psychology and education is the Rasch model, named for the Danish mathematician, Georg Rasch (Rasch, 1966, 1980). Rasch proposed that performance on a test can be predicted from the ability (A) of the examinee and the difficulty (D) of the item. Embretson (1996) praised the advantages of the additive model, based on a fundamental tenet of measurement theory—additive decomposition—in which two parameters are related to a third variable (e.g., a measurement scale) by an additive (subtractive) relationship. Embretson said, "In the Rasch model, additive decomposition is achieved; the log odds that a person endorses or solves an item is the simple difference between his or her trait level….and the item's difficulty…" (p. 348).

The SB5 and Rasch Analysis

The Rasch model was used in several ways and in several stages of the development of Stanford-Binet Intelligence Scale, Fifth Edition (SB5, Roid, 2003a). Some of the important uses of the model and its advantages included item analysis, item calibration, and development of change-sensitive scores (CSS) for each of the major summative scores (4 IQ scores and 5 cognitive-factor indexes). With the Rasch model, both item difficulty and examinee ability are scaled in the same measurement metric. Difficulty calibrations and ability are initially estimated by computer programs and the values appear as normal-curve *z*-scores (called "logits" or log units, Lineacre & Wright, 2000), ranging from minus 4.0 to plus 4.0. For better interpretability, the difficulty values

for each SB5 item and the resulting CSS (estimates of examinee ability) were converted to the W-scale developed by Woodcock and Dahl (1971). The W-scale transforms the initial logit values by centering them at 500 and using a special expansion factor of 9.1024, developed by Woodcock and Dahl. Thus, the CSS scale and item difficulty scale for SB5 ranges from approximately 425 for 2-year old children to 525 for adults, with a central value of 500 located at the mean performance level of children 10 years, 0 months of age (beginning fifth grade approximately). The CSS scale and item difficulty have a criterion-referenced interpretation based on age equivalence, task characteristics (e.g., complexity of the SB5 items), and overall sequence of cognitive development suggested by the scale. As a child progresses upward on the scale, he or she is capable of mastering increasingly complex tasks and solving increasingly challenging problems. This progress mirrors the development of the brain, the growth of academic competencies, and the accumulation of general knowledge. In addition to norm-referencing, where the child is compared to peers of the same age, the CSS scale allows for criterion-referencing to task complexity, and age-related milestones such as the achievement of reading fluency or the various stages in mathematical competence.

CSS scores are available for Full Scale, Nonverbal, Verbal, and Abbreviated IQ and for the five cognitive factors from the Cattell-Horn-Carroll theory (Carroll, 1993; Horn & Cattell, 1966; Flanagan, 2000). When these CSS scores are plotted across age groups, using cross-sectional (not longitudinal data), the classic "growth curve" shapes are evident. The cognitive-factor curves increase from the early childhood years through the early twenties, and, then, depending on the cognitive factor being measured, begin to show declining scores in older age groups. Memory CSS scores show the most rapid decline across elderly age groups, perhaps due to the emergence of dementia, Alzheimer's disease, etc. An exception is the crystallized (General Knowledge and Vocabulary) ability factor which shows continuing improvement into the late 50's among older adults.

Rasch Growth Scores in Other Tests

Previous applications of the Rasch model were made in the Woodcock-Johnson Psychoeducational Battery, Revised (Woodcock & Johnson, 1989), the Toddler and Infant Motor Evaluation test (TIME, Miller & Roid, 1994), in the Leiter International Performance Scale, Revised (Leiter-R, Roid & Miller, 1997), and in the new Merrill-Palmer Developmental Scales, Revised (MP-R, Roid & Sampers, 2004). These instruments and the "growth scores" in them have generally been received positively by professionals working with disabilities or developmental delay. The potential is great for detailed tracking of growth or change across time, and the interpretive power of criterion-referenced scales such as CSS. A striking consistency across national standardizations and across test developers has begun to emerge when the CSS or Growth or W-scale scores are compared across cognitive batteries such as the SB5, the WJ-R, the TIME, the Leiter-R, and the MP-R. The Rasch-based scores on each of these tests have been anchored to the value of mean score of children, age 10 years, 0 months (or, in the case of the MP-R, at 460 for age 4 years, 0 months). Theoretically, the ends of each scale

could them depart in various ways across batteries. However, excellent consistency has been achieved across these diverse test batteries (e.g., consistency of 425 as a value at age 2).

Quality of Performance and Change-Sensitive Measurement

Another important advance in measurement that makes change-sensitive assessment possible is the development of instruments sensitive to the quality of the individual's performance. Rather than simple counts of the number of correct responses or the number of behavioral milestones achieved on schedule (e.g., early vocabulary before age 1, walking at about age 1, learning to read by age 8 or 9), the unique quality of responses can be observed and recorded. Examples of performance quality assessments are listed below and will be described in more depth in the presentation:

Movement Quality in Infants and Toddlers. For example, quality of movement in infants and toddlers was studied as part of the development of a test called the Toddler and Infant Motor Evaluation (TIME, Miller & Roid, 1994). Detailed observations of children with both typical and atypical motor development were taken and detailed illustrations of children in various movement positions were drawn. Examiners using TIME can observe a child moving from a prone position to standing in a 12-month old child, for example. Observations are made every 5 seconds and recorded on the test. The pattern of the movements, not simply the final position (standing) is important in identifying mild and moderate developmental delays. The child should roll over, use hands, arms, and knees to lift himself or herself from the floor, and then use one leg (with perhaps a hand on a chair) to move to a standing position, in the typical pattern. Odd positions of hands, arms, back, legs, etc., may indicate atypical movement. Thus, the quality of the movement is assessed with the TIME system. Miller and Roid (2003) used a sequence comparison method (Jackson, 1990; Sellers, 1974) to compare typical patterns (stored in a computer program) to the patterns observed in typical and atypical children, with excellent discrimination. Details of the method and research will be discussed in the paper.

Quality of Cognitive Performance on the SB5. Guidelines for interpreting the Stanford-Binet, Fifth Edition (Roid, 2003a) include recommendations for the qualitative assessment of child performance on certain subtests and items. For example, the quality of fine motor movement exhibited by children while assembling the pieces of the Form Board or Form Patterns tasks can vary from exceptional, typical, to unusual and atypical movement, modes of grasping the pieces, etc. Most striking, the strategies used by the child to sort the picture chips in the Verbal Fluid Reasoning task are very interesting. The task is to sort the chips into groups of three. Some children only use very concrete categories such as color. Others use functional categories such as "writing utensils," or "play equipment," revealing the quality of their developmental level of thinking. Such qualitative details can be lost if the tasks are not designed to allow their observation or if examiners do not attend to them.

Play-Based Quality of Performance Measures: The MP-R. The new revision of the classic Merrill-Palmer Developmental Scales (Stutsman, 1948; Roid & Sampers, 2004) includes several toy-based tasks that tap the quality of infant and child cognitive and fine-motor abilities. A 'spin toy' reveals the infants quality of hand movement and hand-eye coordination. The 'problem box' (a clear plastic box with interior shelves into which a small toy is inserted with the task to extract the toy) reveals many problem-solving (fluid reasoning) strategies in children. Some children shake the box and pound in on the floor or table. Others try to reach into the small openings in the box. Others discover the bottom "flap" and open it to extract the toy. These toy-based tasks provide great richness of quality performance assessment, and provide indicators of advanced, typical, or delayed/atypical performance for purposes of early identification of developmental disabilities.

Assessment of Essay Writing in School Children. Data on 10,000 students in the public schools of the State of Oregon (USA) were studied by Roid (1994). Essays from these students were graded using a 6-point, analytical trait method of performance assessment with substantial inter-rater reliability. The ratings produce 6 trait scores for each essay (each student) on dimensions such as quality of word choice, grammar and mechanics, creative expression ("voice"), organization, etc. Roid (1994) used cluster analysis to identify groups of students with similar patterns of trait scores and found groups that had high creativity versus poor mechanics of writing.

Assessment of Fluid Reasoning in Infants. One challenging area of assessment is identifying the quality of fluid reasoning in children under the age of 2 years. Prior to work on the Merrill-Palmer revision (MP-R), few published tests provided standardized measures of infant reasoning, except the Bayley Scales of Infant Development and a few others. Also, existing measures did not have "change-sensitive scores" or quality-of-performance items as in the MP-R. Now, the MP-R provides a downward extension of Woodcock's W-scale down to a value of approximately 327 for age 1 month, based on cognitive play-based tasks, observations of eye-movements in tracking toys, etc. These findings will be discussed in the context of the challenge of early assessment of fluid reasoning.

Change Sensitive Assessment and the Evaluation of Cognitive Delays in Premature Infants

Assessing premature infants is an area of important advancement promised by the development of change-sensitive scores and methods of measuring quality of performance. As part of a federally-funded research program, the developers of the Merrill-Palmer Developmental Scale, Revised (MP-R, Roid & Sampers, 2004) have begun to study the problem of using "age corrections" on developmental scales. Because premature infants are often born 4 to 8 weeks prior to typical gestation, scores on their future developmental tests are often "corrected" by using norm tables one or two months lower than the chronological age (measured from birth) for those infants. Lems, Hopkins, & Samsom (1993) suggested that a full correction for children in the first 6 months of life may overestimate the child's score and that a lack of correction will underestimate the child's abilities. The correction may mask a true delay. When, exactly, does the correction diminish and by

what magnitude? Aylward (2002; 1997) suggests that the degree of correction to accurately predict outcomes of premature infants will require an algorithm based on the age of the infant, background risk factors, and, importantly, the domain of cognitive, motor, or language behavior being assessed. Recent research using the new MP-R will be reviewed to show progress made in examining the age correction dilemma.

SUMMARY

More than a decade of research has been conducted to study and develop instruments sensitive to developmental growth and decline in cognitive functioning. Many applications to important assessment problems in education, psychology, medicine, special education, and infant evaluation have been discussed. Many challenges remain for future researchers, including continuing studies using true longitudinal research designs, experimental studies of premature infants "catching up," and studies of early-emerging cognitive abilities such as fluid reasoning. Possible technology developments in the future may be promising, such as use of personal ("palm") data-collection devices to test children more frequently across time. Finally, one of the promising advantages of change-sensitive assessment is the ability to show parents of children with special needs that their children are making progress predicted by the patterns of documented growth curves.

References

Aylward, G.P. (2002). Methodological issues in outcome studies of at-risk infants. Journal of Pediatric Psychology, 27(1), 37-45.

Aylward, G. P. (1997). Conceptual issues in developmental screening and assessment. Journal of Developmental and Behavioral Pediatrics, 18(5), 340-349.

Bauer, D. J., & Estell, D. B. (2001). Cluster analysis of developmental profiles: Relations between trajectories of aggression and popularity over adolescence. In L. Collins & A. Sayer (Eds.), New methods for the analysis of change. (pp. 385-387) Washington, DC: American Psychological Association.

Carroll, J.B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

Collins, L. M., & Sayer, A. G. (Eds.) (2001). New methods for the analysis of change. Washington, DC: American Psychological Association.

Embretson, S.E. (1996). The new rules of measurement. Psychological Assessment, 8, 341-349.

Flanagan, D. (2000). Wechsler-based CHC cross-battery assessment and reading achievement. School Psychology Quarterly, 15, 295-329.

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 57, 253-270.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Jackson, D. F. (1990). The use of sequence analysis algorithms in research on computer-assisted problem solving strategies. Paper presented at the meetings of the American Educational Research Association, Boston, April.

Krishnakumar, A., & Black, M. (2001). Estimating cognitive growth curves from environmental risk factors: Mediating the role of parenting and child factors. In L. Collins & A. Sayer (Eds.), New methods for the analysis of change. (pp. 414-415) Washington, DC: American Psychological Association.

Lems, W., Hopkins, B., & Samson, J. F. (1993). Mental and motor development in preterm infants: the issue of corrected age. Early Human Development, 34, 113-123.

Lincacre, J.M., & Wright, B.D. (2000). WINSTEPS v. 3.00: Rasch item analysis computer program manual. Chicago: MESA Press.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Miller, L.J. & Roid, G.H. (1994) The T.I.M.E. Toddler and Infant Motor Evaluation. San Antonio, TX: The Psychological Corporation.

President's Commission on Excellence in Special Education. (2002). A new era: Revitalizing special education for children and their families. Washington, DC: U. S. Department of Education.

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Roid, G. H. (2003a). Stanford-Binet Intelligence Scales, Fifth Edition examiner's manual. Itasca, IL: Riverside.

Roid, G. H. (2003b). Stanford-Binet Intelligence Scales, Fifth Edition technical manual. Itasca, IL: Riverside.

Roid, G.H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. Applied Measurement in Education, 7, 159-170.

Roid, G.H., & Miller, L.J. (1997). Leiter International Performance Scale--Revised manual. Wood Dale, IL: Stoelting.

Roid, G.H., & Sampers, J. (2004). Merrill-Palmer Developmental Scale—Revised manual. Wood Dale, IL: Stoelting.

Roid, G. H., & Woodcock, R. W. (2000). Uses of Rasch scaling in the measurement of cognitive development and growth. Journal of Outcome Measurement, 4(2), 579-594.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. SIAM Journal of Applied Mathematics, 26, 787-793.

Stutsman, R. (1948). Merrill-Palmer scale of mental tests. Wood Dale, IL: Stoelting.

Torgesen, J. K. (2002). The prevention of reading difficulties. Journal of School Psychology, 40, 7-26.

Woodcock, R. W., & Dahl, M. N. (1971). A common scale for the measurement of person ability and test item difficulty. (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.

Woodcock, R.W., & Johnson, M.B. (1989). Woodcock-Johnson Tests of Cognitive Ability--Revised. Chicago: Riverside.

Wright, B.D. & Stone, M.H. (1979). Best Test Design Chicago: Mesa Press.