Measurement Issues in Evaluating Student Development Programs

Robert A. Mines Counseling Psychology Program, University of Denver

This article introduces the practitioner to psychometric issues related to developmental assessment and makes recommendations for the assessment refinements needed in the field.

Student development practitioners have been challenged to use developmental theories for programming since the early to mid-1970s. In addition to applying developmental models in programming, the student development practitioner needs to be accountable and to use sophisticated evaluation techniques (Kuh, 1979; Mines, Gressard, & Daniels, 1982). To meet these demands, the practitioner must use the techniques of student development assessment.

This study involved an examination of measurement issues related to developmental assessment and the steps necessary to refine the techniques. This article includes descriptions of (a) issues in evaluating developmental stages and tasks, (b) the formats and scoring methods for developmental stage models, and (c) the problems involved in determining the complexity of developmental task models. It also includes recommendations for the development of assessment techniques that could aid practitioners conducting program evaluations.

ISSUES OF DEVELOPMENTAL ASSESSMENT

Assessing development poses problems beyond those encountered in the traditional state-trait, achievement testing, or behavioral assessment perspectives. These issues include:

1. The length of time required for stage or task change versus the duration of many student services programs
2. The complexity and unevenness of the stages across content areas (decalage), which prohibits a global, all-or-none assessment approach

Length of Time for Change

Kitchener (1982) suggested that the length of time needed for cognitive stage change or task resolution almost precludes the possibility that a typical student services program (e.g., freshman orientation, student leadership workshop, or interpersonal problem-solving workshop) would have significant impact on a selected developmental domain. Because development does not end when a student graduates from college, only a specific segment of the developmental process can be measured. As a function of the time required for change, global approaches to developmental assessment usually do not incorporate the sensitivity necessary to encourage change. Methods that assess microdevelopmental changes must be constructed to aid in evaluating programs.

Microdevelopmental changes are those skills or behaviors that represent varying degrees of mastery of a given stage or task. Microdevelopmental changes are assumed to occur in smaller time periods than global stage changes. The methods for appraising these microchanges include the identification of the skills (e.g., identifying better and worse alternatives for solving an interpersonal problem rather than right and wrong alternatives) necessary for more complex reasoning or for resolution of a developmental task (e.g., communicating openly with members of other ethnic or cultural backgrounds).

Stage Complexity and Developmental Decalage

Social cognitive stages or levels have been defined as qualitatively different from other forms of reasoning or thinking because each stage has its own internal logic or set of assumptions about knowledge or reality. These assumptions are used to rationalize problems. Kitchener and King's (1981) reflective judgment levels should be included in this definition. In the social-cognitive stage of development, the assumption is that stages are complex (Rest, 1979). Individuals may exhibit various stage levels in a given context and across content areas, a phenomenon known as developmental decalage. In a complex stage perspective, individuals do not move through the stages in progressive order. For example, in his research on moral reasoning, Rest (1979) demonstrated that on the Defining Issues Test, individuals exhibit varying percentages of reasoning typical of a person at that stage of development, indicating that individuals exhibit a variety of typical developmental stage responses during their reasoning process in more than one stage.

Complexity and decalage of stage phenomena also cause problems in developmental task assessment. The complexity of task assessment is attributable to the variety of basic psychological processes or domains inherent in task resolution (e.g., cognitive, behavioral). The task may be assessed in an all-or-none manner when, in fact, the task resolution is complex. The attainment or resolution of developmental tasks consists of changes in attitudes and reasoning processes as well as behavioral changes. Chickering's (1969) vector of freeing interpersonal relationships is a good example of a developmental task. The assessment of developmental task resolution requires a complex, multilevel approach because of the apparent interaction of cognitive, behavioral, and environmental factors.

The implications of evaluating these program issues are:

1. Because stage change or task resolution occurs over a number of years and in many cases is not complete after 4 years of college, the program evaluation should focus on the identification of the microdevelopmental process or steps that can be taught in a short time.
2. Because the development of reasoning skills may be context specific, the program evaluator should develop the assessment technique for the content area of the intervention. Using a universal, all-purpose measure of development may obscure any true changes that occur.
3. The complexity of the stage or task necessitates the specification of the component of the stage that is being assessed (e.g., evaluation skills in interpersonal relationships, problem solving), the psychological domain (e.g., cognition, behavior) or the use of multiple measures across domains (e.g., freeing of interpersonal relationships, which can be assessed cognitively using the Mines-Jensen Interpersonal Relationship Inventory [Mines, 1978] and behaviorally using the Student Development Task Inventory [Winston, Miller, & Prince, 1979]). This will help the program evaluator better determine changes related to the program. Test format and scoring variations also contribute to interpretation problems in evaluating developmental changes.

FORMATS FOR MEASURING STAGE DEVELOPMENT

The assessment format determines the information obtained. Rest (1976) distinguished betweeen preference, comprehension, and spontaneous use (production) of responses associated with a particular cognitive stage. The most conservative and taxing measure of stage level(s) is to have the individual produce his or her stage process in response to the test items. A production response is elicited through an interview or pencil-and-paper format in which the person is asked to give his or her point of view and a rationale for that point of view on a given problem (similar to oral or written comprehensive examinations). Essentially, persons are asked to demonstrate their reasoning process. Various formats have been used to elicit preference (Likert-type scales), comprehension (asking students to paraphrase or match statements) or production (open-ended or structured interviews) responses. Individuals prefer a higher stage response than they can comprehend or produce. They can comprehend a higher stage response than they can produce. The variety of formats that elicit different levels of stage acquisition limit the meaning of inferences that can be made about stage level development without considering the assessment method used.

Production Formats

The responses of the open-ended interview, semistructured interview, and sentence completion formats are evaluated by comparing their similarity with typical responses at this stage level or with scoring rules. The Reflective Judgment Interview (RJI) (King, 1977) is an example of a production format. The RJI is a semi-structured interview in which the participant reviews his reasoning process aloud. The responses are tape-recorded, transcribed, and then evaluated according to a set of scoring rules. The scoring rules include examples of statements typical of that stage level. Because these evaluations are subjective, they must be rated by two or more persons.

There are two advantages to this technique. An open-ended data source allows the refinement of theory. In addition, the production format provides a rich source of specific stage statements that can be used in objective test development. The development of the Defining Issues Test (DIT) (Rest, 1979) for moral reasoning is a good example of how interview data can be used to develop an objective instrument. Disadvantages of this format are the amount of time needed for data collection and the expense of having the data rated (Mines, 1981). This assessment format is best suited for evaluation of long-term impact, but it is not practical for a typical short-term program evaluation.

Preference and Comprehensive Formats

Preference or comprehensive formats are usually presented in a Likert-type scale or multiple-choice form. These formats are used when the theory and typical stage responses have been identified and the purpose of the assessment is to classify a person's stage level in a systematic manner (Rest, 1976).

The Likert-type scale uses a consistent set of stage-typical statements, which minimizes the chance of receiving an inappropriate response or one that is ambiguous. The objective format, however, does not allow determination of the underlying developmental process used by the student to make the choice and is easily faked. This assessment format is most amenable to descriptive evaluations of the stage levels of the participants in a program. It also has the advantages of being easily administered to groups and relatively inexpensive to score and interpret (e.g., Rest, 1979). It is not suitable for evaluations of outcome for short-term programs because developmental change is slow.

The format of the assessment technique directly affects what the program evaluator can infer about the degree of developmental stage or task consolidation. The format also places practical limits on the size of the sample that can be used in the evaluation. The production formats have generally been used with samples of 125 students or less. A rule of thumb for estimating the time required for a production format is about 2 1/2 hours per student (1 hour for the interview, 1 hour for the transcription, 1/2 hour for rating and coding). The recognition and preference formats permit larger samples because the data can be machine scored (e.g., DIT, Rest, 1976). The larger samples increase the accuracy of the statistical methods used to analyze the data. The probability of finding small but significant changes increases when the accuracy of the statistics increases. The format selected for evaluating developmental change related to programming depends on the complexity of the change, the speed at which change occurs, the intent of the evaluation (i.e., description or outcome), and the specificity of the desired outcome.

SCORING METHODS FOR DEVELOPMENTAL STAGE ASSESSMENT

The scoring method has direct impact on stage classification or description. There are various scoring schemes, each yielding different information. These methods may use either the highest scored stage, the modal level of stage usage (Loevinger, 1976), the mean level of stage usage (Kitchener & King, 1981), the percentage of the highest stage exhibited (Rest, 1976), cutting scores that use cumulative distributions of responses typical of each stage (Loevinger, 1976), or a strong scalogram analysis (Fischer, Hand, & Russell, 1984). In the scalogram technique a task or skill is selected that an individual at a given stage should be able to complete successfully but an individual at the stage or level below should not be able to complete.

The use of the highest stage score assumes that individuals will produce their highest stage responses, which they are not always motivated to do. A potential problem with the use of the highest stage score is assuming a simple stage model intepretation. None of the major theorists (e.g., Fischer, et al., 1984; Loevinger, 1976; Rest, 1976) have assumed a simple stage model. Because development may be uneven and may vary across content domains, the use of a single stage score does not represent the complexity of the phenomena. In addition, an evaluator or researcher may want to know only the lowest level of production (e.g., the stage used under stress).

The use of the modal stage response potentially has the same simple stage assumption problems. The modal stage response also underestimates the highest stage production or comprehension level of the student.

Using the percentage of the highest level produced is a move toward a more precise description of the complex stage properties. The use of only the highest stage percentage ignores the percentage of lower stage responses exhibited.

The average stage score initially takes upper and lower stage scores into consideration but eliminates the stage variance through the use of the mean. For example, the average multiple stage score is used in the Reflective Judgment scoring rules (King, 1977). The complexity of the stage level is diluted, however, by averaging the stage levels across two raters. The diluted stage level becomes a data reduction problem. This results in a conservative estimate of stage functioning that is also affected by motivation and decalage problems with the test items.

Loevinger (1976) addressed the complex stage scoring problem by using ogive rules of cumulative distributions. Ogive rules are cutting scores that use the distribution of responses rather than the mean, median, or mode. Unfortunately, without a breakdown of stage responses by student, it remains difficult for the evaluator to use the results from ogive rules in a sophisticated manner for programming or evaluation because the stage score does not convey the intricacy or the interplay of the different stage skills or stage assumptions.

Fischer et al. (1984) offered a variety of innovative scoring procedures for cognitive development measures. They suggested that a strong scalogram could be used. This procedure predicts a sequence of steps in acquiring developmental skills within a specific content domain. A separate task is designed to assess each step. Each individual's performance should fit a Guttman Scale. This method eliminates the scoring rule problems discussed previously and also eliminates the problem of using one task to differentiate all developmental stages of a model. "When every developmental stage is assessed independently, the assumption (the use of a single developmental task) is no longer a problem since it becomes a hypothesis to be tested" (Fischer et al., 1984). To date, such an approach to student development assessment does not exist in a format that is useful for student services practitioners.

The assessment format and scoring rules result in a wide range and variety of stage scores. The range and variety are attributable to the inherent problems of decalage, the qualitative differences of each stage, and the complexity of stage change. The assessment of developmental tasks has a related yet different set of psychometric problems.

DEVELOPMENTAL TASK MODELS

The developmental task models differ from the stage models in the conceptualization of adult development. The developmental tasks are culturally specific and occur at approximately the same time in the life of a given age group. The task must be successfully completed to provide the experiential foundation needed to resolve later developmental tasks. If a developmental task is not resolved, theoretically the student will not have the foundation for the successful resolution of subsequent tasks.

Cognitive Complexity and Developmental Tasks

The developmental task models present a complex measurement problem and a specific goal for young adults to attain (e.g., Chickering's [1969] freeing of interpersonal relationships). The attainment of this goal requires multilevel (e.g., cognitive, process skills, behavior) changes. On one level, from a theoretical standpoint, changes in cognitive complexity should occur. For example, in his vector of freeing interpersonal relationships, Chickering (1969) assumed that movement from dependence to independence to interdependence (as well as increasing one's tolerance for diversity) indicated the individual experienced a shift in cognitive complexity. An increase in cognitive complexity, as described by Perry (1970) or Kitchener and King (1981), is theoretically necessary to move from banal, stereotypical views of others to a sophisticated appreciation of individuals from diverse backgrounds. An increase in complexity is implied by being able to process interrelationship issues and the relevant compromises and benefits necessary to relate interdependently without becoming dependent or counterdependent.

Also implicit in the resolution of this task are changes in ego and moral development stage levels. The ego development changes are necessary for an interdependent self-awareness. The moral development changes reflect the implicit social contracts of dependent relationships versus interdependent relationships. Thus, the assessment of cognitive stage change or attitudinal change is one aspect of evaluating a development task.

Skills Related to Developmental Tasks

The developmental tasks can also be considered from the perspective of the skills needed to complete the task. For each task, the individual must exhibit certain skills to resolve the task. For example, the freeing interpersonal relationships vector requires the individual to exhibit certain communications skills to function interdependently. Assertiveness skills, conflict mediation skills, and basic communications skills, such as those mentioned by Egan (1982), would be helpful for living interdependently. Data does not exist, however, regarding specific skills required for task resolution. As with the cognitive stage assessment problems, the identification and assessment of task-related skills may provide a more refined understanding of the components of developmental task resolution. These task-related skills may be the most promising area for identification, assessment, and intervention in a developmental framework.

Task-Specific Behaviors

In addition to cognitive and process skills, developmental tasks can be characterized by task-specific behaviors. The assumption is: If the student exhibits task-related behavior, the student has resolved the task. This perspective has some appeal because it eliminates the problem of assessing cognitive stage, skills, and attitudes and deals directly with the behavioral manifestations. This approach is probably sufficient if the behaviors can be identified and the evaluator is only concerned with a yes or no diagnosis of task resolution. Unfortunately, the resolution of a developmental task is probably a process, not an event. The process occurs over time and, as noted previously, involves cognitive and skill changes as well as behavioral changes. Assessing only behavior tells the student services evaluator nothing about the process fundamental to an individual's progress in task resolution or about programming needs related to task resolution.

The best approach to developmental task assessment integrates the cognitive stage, skills, and behavioral dimensions. The second best alternative identifies specific skills related to task resolution as well as behavioral manifestations of the task. The least desirable approach assesses only one dimension or one aspect of the dimension. Data are not available to delineate the relationships of cognitive complexity or attitudinal changes to behavior indicators of task resolution. Relatively little information on development of a developmental task instrument is available to the public (with the exception of the work of Winston et al., 1979) that would make the evaluator's instrument selection easier.

The evaluator using a developmental task approach needs to be aware of the issues discussed above. With these issues in mind, the evaluator can identify specific behaviors, skills, or cognitive processes that should at least have a theoretical relationship to task resolution in a given program. The behaviors, skills, or cognitive processes can then be assessed by homemade, self-report instruments or observations of actual task-related behaviors. The creation of these homemade systems can be based on the evaluator's skills in setting a goal and developing performance objectives for the management arena. Creativity will be particularly important for evaluating short-term program outcome because the evaluation techniques will vary widely, depending on the content area of the program. In addition, multilevel assessment will be needed for more sophisticated outcome evaluations that increase the degree of certainty of the conclusions.

The developmental task approach is appealing for specific programs that can enhance task resolution, such as assertiveness training or cross-cultural training for freeing relationships. As more is learned about specific skills related to task resolution, assessment techniques can be developed that will relate to those skills, and these techniques can take the place of the global approach presently used.

CONCLUSION AND FUTURE DIRECTIONS

The assessment of development stages and tasks of young adults has not been sufficiently refined to allow student services practitioners easy use of the techniques for short-term program evaluation. This places the practitioner in an untenable position between the demands for accountability and the existence of few standardized means to achieve it.

New assessment techniques and methodologies are needed. One direction these techniques may take is the incorporation of complex stage scoring systems that will allow more meaningful interpretations to be made in long-term evaluations. A second direction would be the use of multilevel assessment for developmental tasks, which would tell the evaluator which aspect of a task has been affected by a program. The third and perhaps more innovative and practical direction evolved from the work of Fischer et al. (1984), who argued for the implementation of a strong scalogram analysis. In their construct, the student would practice the developmental skills and tasks both before and after the program to determine maximum developmental performance. Thus, for the evaluation of the typical program outcome, the specific skills related to a given task or stage (microdevelopmental sequences) must be identified and techniques must be developed to assess the skill. This assessment involves either skill-specific demonstrations or in vivo observations.

Chickering's (1969) vector of freeing of interpersonal relationships has a skills component that may be amenable to a strong scalogram procedure. An example of a developmentally based skill is problem solving, which may be necessary to live interdependently. This skill can be assessed by a pencil-and-paper format such as the Means-End Problem Solving Procedure (Platt & Spivak, 1975) or actual observations of the use of problem-solving skills in role playing or live conflict resolution during the program. The effectiveness of the program can then be demonstrated without encountering the problems associated with the global techniques or with the slowness of overall change.

The new class of instruments should incorporate specific skills or processes related to a given stage rather than global descriptors. These instruments must be cost effective, objectively scored, and easily administered in group situations to be useful for the student services practitioner faced with the challenge of developmental programming and evaluation.

REFERENCES

Chickering, A. (1969). *Education and identity.* San Francisco: Jossey-Bass.

Egan, G. (1982). *The skilled helper.* Monterey, CA: Brooks/Cole.

Fischer, K. W., Hand, H. H., & Russell, S. L. (1984). The development of abstractions in adolescence and adulthood. In M. Commons (Ed.), *Post-formal thought.* New York: Praeger.

King, P. M. (1977). *The development of reflective judgment and formal operational thinking in adolescents and young adults.* Unpublished doctoral dissertation, University of Minnesota.

Kitchener, K. S., & King, P. M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. *Journal of Applied Developmental Psychology, 2,* 89–116.

Kuh, G. D. (1979). *Evaluation in student affairs.* Cincinnati: ACPA Media.

Loevinger, J. (1976). *Ego development: Conceptions and theories.* San Francisco: Jossey-Bass.

Mines, R. A. (1978). *The Mines-Jensen interpersonal relationship inventory.* Paper presented at the meeting of the American College Personnel Association, Detroit.

Mines, R. A. (1981, May). *Psychometric aspects of the reflective judgment interview procedures.* Paper presented at the meeting of the American College Personnel Association, Cincinnati.

Mines, R. A., Gressard, C. F., & Daniels, H. (1982). Evaluation in student services: A metamodel. *Journal of College Student Personnel, 23,* 195–201.

Perry, W. G., Jr. (1970). *Forms of intellectual and ethical development in the college years.* New York: Holt, Rinehart and Winston.

Platt, J. J., & Spivak, G. (1975). *Manual for the means-end problem-solving procedure (MEPS).* Philadelphia: Hahnemann University, Department of Mental Health Services.

Rest, J. R. (1976). New approaches in the assessment of moral development. In T. Lickona (Ed.), *Moral development and behavior.* New York: Holt, Rinehart and Winston.

Rest, J. R. (1979). *Revised manual for the defining issues test.* Minneapolis: Minnesota Moral Research Projects.

Winston, R. B., Jr., Miller, T. K., & Prince, J. S. (1979). *Assessing student development.* Athens, GA: Student Development Associates.