DOCUMENT RESUME

ED 340 765 TM 018 021

AUTHOR Rock, Donald A.
TITLE Development of a Process To Assess Higher Order Thinking Skills for College Graduates.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Nov 91
NOTE 40p.; Commissioned paper prepared for a workshop on Assessing Higher Order Thinking & Communication Skills in College Graduates (Washington, DC, November 17-19, 1991), in support of National Education Goal V, Objective 5. For other workshop papers, see TM 018 009-024.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adult Literacy; Cognitive Measurement; *College Graduates; Communication Skills; Critical Thinking; *Databases; *Educational Assessment; Higher Education; Models; *National Programs; National Surveys; Problem Solving; *Program Development; Scoring; Test Construction; *Thinking Skills
IDENTIFIERS National Education Goals 1990; *National Information Systems

ABSTRACT

Issues in the development of assessments of higher order thinking skills for college graduates are discussed in the order in which they were presented when this series of papers was commissioned. With regard to Issue 1, it is generally agreed that the development of these skills is a desirable goal, but there is little consensus on how they should be assessed and less agreement on how they should be taught or measured. The next step in development of large-scale assessments will be the use of scoring protocols that are developed to provide diagnostic information for instruction. The development of extended free response items for large-scale assessments is discussed, with reference to existing national surveys such as the National Assessment of Educational Progress and the National Adult Literacy Survey. The setting of performance standards as posed in Issue 2 is discussed. The only currently relevant database, as questioned by Issue 3, is the Graduate Record Examination database assembled for the National Science Foundation, which could serve as a beginning and a model for development of national assessment data. A six-item list of references is included. Reviews by L. Boehm, J. L. Herman, and M. Scriven of this position paper are provided. (SLD)

DEVELOPMENT OF A PROCESS TO ASSESS HIGHER ORDER THINKING SKILLS FOR COLLEGE GRADUATES

Donald A. Rock

Educational Testing Service

Issue One

As requested in the "guidelines" I will discuss the issues in the same order as they were presented in the guidelines. The guidelines in issue one ask: "What specific skills would or should be affected by cognitive and/or affective learning experiences, as they relate to critical thinking, communication and problem solving abilities?" Issue one goes on to say that-

> " there should be a common understanding and perhaps agreement, about what specific skills students should achieve considering the larger goal of developing a competitive workforce and enhancing citizenship skills. Once clarified, they must then be defined in a manner to allow for measuring or assessing the impact of the educational experience. In addition they should be defined from a teaching/learning perspective, so that their enhancement can be factored into classroom experiences".

Issue one implies a closer tie than probably exists between the skills that are required in the workplace and those that are taught in college. However there appears to be some agreement among both academicians and employers about the desirability of developing one's problems solving skills, critical thinking skills, and communication skills. While it is generally agreed that the development of these skills is a desirable goal there seems to be very little consensus on how they should be assessed and even less agreement on how they should be taught.

The phrase -"they should be defined from a teaching/learning perspective, so that their enhancement can be factored into classroom experiences"- suggests that we must not only define these skills in such a way that assessment tasks can be constructed, but that their definition be sufficiently detailed that the appropriate teaching behaviors immediately follow from the definition. If such definitions can be formulated and agreed upon, then the relationship between assessment and instruction will become much closer than it has traditionally been in the past.

This tying of assessment to instruction would seem at "first blush" to go well beyond the intent of the past national assessments of cognitive skills e.g., NAEP, NELS, and NALS. However, the 1992 NAEP assessment is in some content areas moving closer in the direction of merging assessment and instruction. Until very recently NAEP defined through its test specifications and associated item pools what students can or cannot do. Such assessments provide a good barometer of educational progress along those dimensions

defined by the item tasks, but they say little about how any identified educational deficits can be remedied.

To a certain extent this has been the history of large scale assessment because the cost of building assessment instruments that not only provide educational indicator type information but also diagnostic information of sufficient depth and scope to inform teaching practices has in the past been simply too expensive.

In addition the test constructors have tended to shy away from building cognitive tasks which would influence the direction of the curriculum. There seems to have been kind of an implicit understanding between the curriculum people, educators, and test constructors that the tests would provide educational status indicators but would leave the decision on how to remedy the educational deficits that are identified to the educators.

NAEP more recently has made the decision to go a little further along the road towards providing feedback to the policy makers by presenting data on not only what students know and can do but also the extent of the gap between what they know and what they **should** know. National standards were set with respect to what students should know by the National Assessment Governing Board (NAGB) and applied to the NAEP scales (Bourque & Garrison, 1991). While there is strong sentiment for setting standards or goals on assessment instruments , this is still only a small step in the direction of providing

diagnostic information that can inform instruction. More will be said about goal setting and its relevance for assessing higher order thinking skills in college graduates.

NAEP is continually increasing the proportion of extended constructed (free) response items in its assessments. Such items have the potential of tapping the kinds of skills referred to above and if properly scored can provide diagnostic information for instructional purposes. The term "extended" free response items as opposed to simply free response items refers to a written highly developed response to a prompt rather than simply filling in a blank with a single word, responding with a single calculation, or a very brief written answer. In many cases the extended free response requires the respondent to show all steps in the solution. At present most of the extended response items are only scored as right or wrong and thus limiting their potential usefulness as diagnostic tools. Even though optimal use is not yet being made of extended free response items, a large data base is being developed, particularly for 12th graders which can serve as an extremely valuable research tool in the development of assessment tasks to be used for college assessment.

The next step in the evolutionary development of large scale assessments will be the use of scoring protocols that are developed specifically to provide diagnostic information for instruction. As indicated above NAEP and to a lesser extent the National Educational Longitudinal Study (NELS:88) and The National Adult Literacy Survey (NALS) are laying the groundwork for the development for scoring systems which go well beyond just providing indicator information. Because the assessment community is moving in the direction of

extended free response formats, NAEP psychometricians are developing partial credit scaling models that are more appropriate for the complex scoring protocols that will be used in NAEP and NELS.

NELS:88 is a longitudinal study that began with a probability sample of 8th graders in the Spring of 1988 and who were subsequently followed up and retested in the 10th grade and eventually (1992) will be retested once more in the 12th grade. NALS is a nationally representative survey of households which includes a literacy assessment. All of these studies NAEP, NELS:88, and NALS originated at the National Center of Education Statistics (NCES) of the Department of Education and are being monitored by the same agency.

The trick here is to be able to develop a scoring system which can be done quickly and cheaply yet provide diagnostic information. Clearly the non-extended free response items as used in both NAEP and NALS can be relatively cheaply scored but their associated prompts are not designed to yield much in the way of diagnostic information. This is not to say that the non-extended free response as presently used in NALS does not offer certain advantages over the classical multiple choice item. The non-extended response format minimizes guessing which has always been a problem in the multiple choice format.

It should be noted that extended free response items can be either holistically scored and/or analytically scored. The term holistic scoring typically assumes that the skill being measured is essentially unidimensional and individuals can be rank ordered along this single

dimension. The amount of diagnostic information available from holistic scoring is limited unless the scoring levels associated with a particular task are specifically defined in terms of developmental diagnostic categories. For example, in mathematics a particular prompt could either explicitly or implicitly require the respondent in his response to demonstrate: (1) a basic understanding of the concept involved in the problem, (2) sufficient knowledge to carry out a logical sequence of operations based on the conceptual understanding exhibited in step (1), and (3) generalize the results, e.g., write and explain a general equation which describes the application. Tasks such as this require the respondent to carry out a sequence of problem solving steps where success at each succeeding step requires knowledge from the preceding step plus additional new knowledge or understanding.

A class of extended free response items that has multiple prompts where each prompt is associated with a level in the knowledge hierarchy will be referred to as a hierarchical structured free response item. An important difference between these structured free response formats and the typical completely open ended free response item is that there is a separate prompt for each of the three levels of problem solving while the typical extended free response item has only one prompt and it typically corresponds to the highest problem solving level in the more structured free response item.

Free response tasks such as this are scheduled to be administered to high school seniors in 1992 as part of the NELS study. Many of the above type of free response tasks have been developed by Romberg and his associates (1991) and modified for use in the

NELS study. The scoring protocols associated with this type of item are relatively easy to implement compared to analytical scoring procedures where the task is inherently multi-dimensional. Rock et. al.(1990) have used multiple choice item patterns to achieve the same kind of diagnostic information. That is they analyze response patterns to sets of items associated with each level in a cognitive demand hierarchy. They refer to a particular test having these hierarchical sets of items as an hierarchically ordered skills test (HOST). Romberg refers to free response items that follow this hierarchical framework as "super" items.

These structured free response "super" items appear to be particularly appropriate for those assessments where there is little reason for the respondent to be highly motivated to perform. In large scale assessments such as NAEP there is considerable concern about the lack of motivation among the respondents. Since at present, there is no accountability at the individual, school, or district level, there is little reason for an individual to do his or her best. Analysis of NAEP and NELS responses suggest that the number of omits and items not reached increases proportionately to the number of free response items present. Analysis of the 1990 NELS field test suggests that one gets better cooperation if one paces the respondent through the free response items. In addition, the more structured free response items as discussed above are more likely to elicit responses at least for the first cognitive demand level then is the typical unstructured free response item with only one prompt.

When an individual does not respond at all to an extended free response item, which is more likely to happen in the extended unstructured format, the item yields no diagnostic information. For example, we do not know if the individual did not understand the task, i.e., has not achieved cognitive level (1) or he or she did not respond because of the higher order skills required to correctly respond to any one of the succeeding levels in the hierarchy, or worse lacks sufficient motivation to go through the effort to construct a response. The structured free response gives even low achieving individuals the opportunity to respond since the first prompt in the hierarchy generally only asks about basic information concerning the problem being posed. Since the structured free response are less likely to lead to blank responses we can then also ascertain whether or not the individual attempted the item.

It is also argued that the structured free response item is more easily adapted to standard setting procedures. In fact, the ordered levels associated with each of the prompts could be used to define various criterion referenced levels.

Preliminary analysis of NAEP results with non-structured extended free response items also suggests that when multiple choice items follow the free response items, they are also less likely to be attempted. To make matters worse there is preliminary evidence from both NAEP and NELS that the propensity to omit or not attempt extended free response items differs by ethnic/race groups. That is, there seems to be a proportionately greater tendency for black and Hispanic students to omit or not reach (i.e., not attempt) extended free response items then is the case with multiple choice items. It is possible that when

minority group students take multiple choice items they are more likely to guess than leave the item blank. Additional research needs to be done in this area before one devotes an entire assessment on a national scale to extended free response. Fortunately much of this research is presently being carried out in both NAEP and NELS.

Additional evidence for the possibility of the lack of cooperation and its impact on interpretation of the results that may occur in a non-risk assessment situation comes from a recent field test of the NALS items using non-extended free response items. Adult subjects ages 16-75 were randomly selected from a household sample and randomly assigned to one of three treatment conditions. Individuals in the first condition received no compensation for performing the testing task. Individuals in the second and third condition received twenty and thirty dollars respectively. Individuals in the zero dollar condition were less likely to agree to do the task then those in either of the remaining two paid conditions. In addition, the zero dollar test takers had a significantly higher mean score than those in the other paid conditions. Individuals who agreed to taking the test in the non-paid condition compared to the paid conditions tended to be better educated and less likely to be a member of a minority group. It would seem that the less academically able are less likely to offer to demonstrate their lack of knowledge or skills without some incentive.

Without some external motivation, in this case financial incentives, one would get a biassed estimate of the population literacy rates. That is, without some type of incentive,

one might lose the cooperation of some of the less able individuals and thus arriving at an overestimate of the population literacy rates.

Interestingly enough once an individual agreed to take the test there did not seem to be any relationship between incentive level and number of omits or number of items not reached. It should be kept in mind that these are non-extended free response items and deal with reading tasks that adults are often faced with in their day to day activities. The point here is that college students may or may not behave as the individuals in the household sample, but we need more information on their behavior in a non-risk assessment situation.

The NALS full scale assessment in the spring of 1992 may cast some light on the cooperation levels of college students versus other household members of the same and different ages. There will be about 25-26000 individuals in this household survey. It is anticipated that 20-25 percent will be in college. Of the 5-6000 individuals who are in college approximately 1500 will be college seniors. Subsequent analysis of the response rates of the 1992 NALS college seniors will provide some "hard" information about their propensity to cooperate in taking a literacy test in a non-risk situation.

Some of ETS's experience with college level assessment programs is also relevant here. ETS has developed or is doing research on a number of college level assessment instruments. One such program is the Academic Profile instrument. One of the purposes

of the Academic Profile is to measure student growth at both the institutional and the individual level. There are two forms varying in length depending on whether the scores are to be reported at the individual or school level. The long form takes about 2 1/2 hours and the short form about 40 minutes. Both norm referenced and criterion referenced scores are reported. Norm referenced scale scores are reported in humanities, social sciences, natural sciences, college level reading, college level writing, critical thinking, and mathematics. Criterion referenced scores which are reported as proficiency levels are given in writing, reading/critical thinking, and mathematics. There are three pre-defined proficiency levels within writing, reading/critical thinking, and mathematics.

When the instrument is used to measure growth or "value added" at the institutional level, the distribution in terms of percentages above each of the three proficiency levels is reported for freshman and then as juniors or seniors. This type of reporting (i.e., distributions above various pre-defined scale points) is very similar to the approaches taken in NAEP and NELS. NAEP has been reporting distributions of percentages above behaviorally anchored scale points and the their changes over time. More recently NAEP in the trial state assessment reported distributions in terms of percentages above selected scale points that were based on NAGB's goal setting procedures.

The norm referenced reporting in Academic Profiles are based on groupings by the Carnegie classification and are provided by class e.g., freshman, sophomore, and juniors or seniors. It is the opinion of the present author that both types of scores are necessary. The

criterion referenced information provides at least some diagnostic information, but its usefulness is enhanced by some appropriate frame of reference i.e., appropriate norm group as is done in the Academic Profiles.

One of the primary drawbacks of the Academic Profile is that it relies on multiple choice items except for a supplementary essay writing task. This of course, leads to relatively speedy reporting with relatively little cost and more completed tasks. The negative here is that little information of a diagnostic information is being reported here.

In summary, while extended free response items seem to promise more diagnostic information, they probably will fall far short of expectations unless the respondent can be motivated to attempt these types of items that require more effort since the student must construct a response. It is the position of this author that unless the respondent can be externally motivated through assignment of accountability, the proposed assessment of the skills of college students should be carried out with a mix of multiple choice and structured free response. The structured free response being used for only those areas that clearly are inappropriate for multiple choice or other objective methods. The structured free response item development should lean on some of the theoretical work on cognitive hierarchies as described by cognitive psychologists such as Maier (1986). Much of the work of Rock & Pollack with HOST tests uses a similar conceptual framework to Maier. Problem solving items using a hierarchical structure can be made relatively general and thus form an item construction paradigm that could be appropriate for a number of content areas. The

structured free response should be paced unless some sort of external motivation can be injected into the testing situation. Further research needs to be done on the differential propensity to attempt extended free response items versus structured free response items by racial/ethnic groups and its relationship to both ability and level of risk associated with the assessment situation.

Issue Two

Issue 2 is concerned with whether performance standards should be set and how. The answer from my perspective is a qualified yes. The answer is yes because:

(1) Setting realistic goals should be the first step in bringing about change in any complex delivery system. The goals must be realistic or the resulting frustration on the part of both teacher and student will only make the situation worse.

(2) As indicated above the setting of performance standards in assessment tasks may help with motivational problems.

(3) Performance standards can provide some diagnostic information concerning educational deficits, and if they (the standards) are accompanied by specific behavioral descriptions they could give some direction as to how to remedy those deficits.

As indicated above the National Association Governing Board set standards for the 1990 NAEP mathematics assessment for the 4th, 8th and 12th grades. Much can be learned from this ambitious effort about the difficulties involved in setting standards. One thing that was learned from the NAGB experience was that if you are to attempt a consensus goal setting procedure where the consensus must be across a set of individuals from very different backgrounds and training, very specific descriptions of the types of behaviors that are to be associated with each standard level must be presented before the judging task can begin. The more diverse the judges the more structured the definitions must be. NAGB in the very first trials attempted to arrive at a consensus across a broad occupational cross-section of "experts" while providing little in terms of behavioral definitions associated with various performance levels. Later "rounds" were characterized by more structured descriptions of the desired behaviors.

If the experts were teachers working at the grade level at which they were setting standards then less structure would be needed. However, to get the support of the general public in addition to that of the educational community, it is necessary to include in the standard setting procedure the consumers of the educational product (representatives of industry and government) as well as the producers, i.e., the teachers themselves. The charge to the goal setters in the NAGB standard setting exercise was to conceptualize what students **should** know with little emphasis on what they **do** know. Without considerable additional information about what students are presently taught as well as how students actually do perform this can be an invitation to set unrealistically high standards.

In all fairness to NAGB the expert judges were given access to item performance information for the grade level they were judging. However, there would have been better articulation between the grade levels on the three performance standards if the judges were shown the item performance for all three grade levels regardless of the grade that they were doing their judging on. I know that there will be those criterion referencing "purists" that will argue that the expert judgements with respect to item performance at each level should be carried out without the help of any empirical data on item performance. This procedure may be defensible when the expert judges are quite familiar with the student performance in question as well as the imperfections of test items as measures of the desired behaviors. This is not the case when the expert judges come from many professions with little experience or contact with formal testing or what is being taught in the classroom.

I would suggest a somewhat different approach to goal setting in a college level assessment than that carried out by NAGB. This is not to say that NAGB did not do a credible job given their relatively severe time constraints. I think that while the consumers of the educational product (industry and government) are reasonably familiar with what skills are required to do jobs in the labor market, they may not be as aware as they should be concerning the less than perfect relationship between performance on a given item and knowledge of a given desirable skill. Even when a particular item appears on the surface to be a rather straight forward measure of a specific desirable skill, it is quite possible that certain overly clever distractors (in the case of multiple choice items) may entice certain individuals who had relatively complete knowledge of the skill involved to get the item

wrong. Professional judges from industry, being less experienced with the foibles of testing, are likely to overlook this tenuous relationship between item performance and the knowledge of the skill in question. As a result they are likely to overestimate the performance of students at a given performance level. Teachers, and other educational specialists are familiar with this phenomena and are likely to take these imperfections in consideration when they make informed judgements on the performance of students at various knowledge levels.

It is suggested that "expert" judges from many different perspectives be used in the goal setting procedure but be given different tasks depending on their expertise. For example, all chosen experts regardless of background should be involved in defining the KSA's (knowledges, skills, and abilities) that are necessary to perform successfully at each of the performance levels in each of the content areas. It is argued here that individuals from industry can play a significant role here since they should be aware of both the level and kind of ksa's necessary to successfully make the transition from college to successful performance in their jobs. Educators and test specialists can then develop and/or match presently available items to the list of desired skills and skill levels.

Then the representatives from industry and government can rate the items in the pool with respect to their relevance to the desired skills. Then using those items that have been judged to be relevant, college teachers who have been involved in the whole ksa process and are familiar with both typical college student performance as well as testing can make the

judgements about expected student item performance for the various criterion referenced levels. This strategy is consistent with the procedures used by industrial psychologists in developing tests to measure skills required in industry.

Issue Three

Issue three is concerned with a number of related practical issues. One question is whether or not the type of information desired is available in current data banks and if not are their proxies available. The answer is not really. Probably the closest thing is a data base recently put together by Educational Testing Service for NSF which includes a longitudinal sample of students who took the SAT and approximately four years later took the GRE's. The data base includes well over a half-million students and about 300-400 colleges. In addition to the test information some biographical information is available on the student as well as considerable information on the characteristics of the college attended.

The problem(s) with this data base are: (1) that it is not a probability sample of either students or colleges, (2) the GRE test is not designed to be a measure of college outcomes, and (3) the multiple choice format and associated scoring procedure of the GRE allow little in the way of diagnostic information. Many of the students also took an area achievement test but these are relatively narrow measures of outcome and the sample at the school level would be quite sparse in any one given content area. However, that being said the major pluses of the data base are:

(1) The availability of the SAT scores as a control variable, and thus growth controlling for input can be estimated.

(2) The fact that an analytic reasoning score is reported in addition to the usual verbal and quantitative scores. While the analytic items are in multiple choice format they can be considered a somewhat limited proxy for critical thinking. The analytical reasoning score can be further broken down into two additional subscores- logical reasoning and analytical reasoning. The logical reasoning attempts to assess reasoning within a more verbal framework while analytic reasoning is more related to quantitative reasoning. It would seem that these two different frames of references could provide limited but reasonably appropriate outcome measures for individuals majoring in the sciences versus the humanities.

(3) The students taking these measures can be assumed to be highly motivated to do their best.

This data base could be considered as a short run proxy for the real thing. Its desirability could be increased if those students who took the GRE and the ACT four years earlier could be merged with those with the SAT and GRE. While not optimal SAT and ACT scores have been equated in the past. This would both increase the size of the data base as well as make it more representative.

The NALS study discussed above could be enlarged to include sufficient household samples to get a respectable sample of college seniors. As it is designed now, if the household individual selected is a college student away from home he or she will be tested. Possibly households could be oversampled if they have one or more family members attending college. The present sampling plan would have to be considerably expanded if one wished to include sufficient college seniors to carry out any relatively detailed analysis even when aggregating colleges by types. As suggested above, there would also be some question of the of how to motivate the college student segment to take the test unless they were paid.

The proposed NALS literacy forms for 1992 are likely to suffer some ceiling effects for some college seniors. Additional more difficult forms could be constructed and linked to the other household forms using common item equating. Even so, it is doubtful that the areas assessed- Prose, Quantitative, and Document literacy are sufficient for the present undertaking. Finally the NALS assessment is not presently set up to provide any in depth diagnostic information. However, more complex scoring protocols could be developed for the present open ended items which could yield a significant amount of diagnostic information. It is my suggestion that the more complex scoring protocols be developed and implemented on the 1992 NALS subsample of college seniors and thus providing relatively immediate feedback concerning literacy performance including diagnostic information on college seniors. The NALS study would also provide at the aggregate level the opportunity to compare non-college going youth of the same age to the college going individuals. This

would give some up to date information on the "value added" component of attending college.

The Academic Profile data base discussed above is a relatively small data base with an over representation of the smaller liberal arts colleges. Unlike the GRE data base, one has to be somewhat concerned about whether or not the students taking the academic profiles are maximally motivated.

I believe that the National Center for Educational Statistics (NCES) is currently planning to collect data on a post-secondary cohort attending a representative sample of post secondary educational institutions. It would seem reasonable to search present SAT and ACT files for test score information on these individuals and eventually administer an outcome measure as described above.

In summary, the only relevant currently available data base is the GRE data base assembled for NSF. As indicated above this data base has many shortcomings but could provide considerable information at a minimal cost while plans could continue to be formulated for developing more appropriate samples and instrumentation.

Issue three also asks who should be tested and with what items and for how long? The answer to all of these questions depends to a certain extent on the accountability issue. If one is modest in their ambitions (at least at first) and simply wishes to assess what college

students know at the aggregate level then the NAEP assessment model would seem to be most appropriate. That is, a sampling of institutions, students within institutions, and a item spiralling design would be most appropriate. Such designs, however, by their nature do not guarantee students performing at maximum levels of motivation unless some external incentive is applied. Such designs also can not be easily adapted to the concerns about relating process to outcomes (e.g., Pintrich, 1986) accept at the most aggregated levels. Relating process to outcomes at the aggregate level may be even less informative at the college level than at the primary and secondary education levels. That is, it is this writers opinion that the higher one goes in the educational system the greater is the opportunity to be exposed to a more diverse set of curriculum and teaching practices. Much of this potential "richness" in process could be lost when aggregation takes place.

The sparse data matrix that results from the spiralling of items can yield quite accurate aggregate statistics such as group means, but does not generally provide optimum estimates of individual scores. The individual level scores are most appropriate for the type of correlational data that is useful in estimating the relationship between process and outcome. The spiralled type of designs, however, provides the opportunity for greater coverage of the ability domain for a given unit of testing time. That is, because of the systematic spiralling a particular student is only assessed on a relatively small part of the total ability domain and thus the testing burden per individual is relatively light.

It is this writers suggestion that NCES consider a two or three phase approach. The first phase would be to take advantage of present data bases such as the GRE and NALS with the suggested modifications outlined above. The next phase would include designing a NAEP like study to get an aggregate picture of college level performance on the skills deemed to be important by representatives of both industry and the educational community. Both the selection of skills and the standards could follow the process outlined in the above session on the second issue. This type of design would allow the gathering (at the aggregate level) the maximum amount of information for the least burden. It will not, of course, be optimal for individual level process-outcome type of analysis, but some process-outcome relationships can be estimated for various aggregations. The spiralled design could also serve as a check on the validity of the conclusions based on the "stop-gap" use of readily available data such as GRE & NALS.

A third phase might include the development of a computer assisted adaptive test (CAT) battery which in theory (Lord, 1981) would allow one to get accurate individual performance estimates for a minimum amount of testing time across a relatively broad set of skills at the individuals convenience. Colleges for the most part have all the necessary hardware to carry out such an endeavor if the software is furnished. Since an individual could be assessed at his or her own convenience at any one of a number of available terminals, cooperation and possibly motivation will be increased. Score reporting would be almost instantaneous since the responses are scored immediately on location and can then be sent to a central location. The necessary item parameters for building the CAT battery

could be gotten from the second phase. The drawback of such a system would be that present technology would force us to rely heavily on the multiple choice framework. Certainly the multiple choice part of the assessment could be carried out this way. The free response part of the assessment could also be carried out on the computer and scored later.

References

Bourque, M. L. & Garrison, H. H. (1991). The levels of mathematics achievement, Vol. 1, NAGB report: Washington, D. C.

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

Maier, R. E., Larkin, J. H. & Kadane, J. (1984). A cognitive analysis of mathematical problem solving ability. In R. Sternberg (Ed.) *Advances in the Psychology of Intelligence*, Vol. 2, pp. 231-273. Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ.

Pintrich, Paul R. (1986). Assessing student progress in college: a process oriented approach to assessment of student learning in postsecondary settings; in Postsecondary Assessment Conference: Report of the Planning Committee, U. S. Dept of Education. Washington, DC.

Rock, D. A., Pollack, J., Owings, J., & Hafner, A. (1990). Psychometric report for the NELS:88 base year test battery. Data series: NELS:88-88-1.4. U. S. Dept of Education, Washington, DC.

Romberg, T. & Colis K. (1991). Assessment of mathematical performance: an analysis of open ended test items. In Wittrock, M. & Baker, E. (Ed) *Testing and Cognition*.

A Review of Donald A. Rock's "Development of a Process to Assess Higher Order Thinking Skills for College Graduates"

Lorenz Boehm

Oakton Community College

I.

A tough essay to review. Dr. Rock assumes an audience that is comfortably housed somewhere near the down-town area of what he refers to as "the assessment community." As a consequence, save for a few notable exceptions, the piece is packed with unclarified jargon and references which, on occasion, leave this reader mystified. In fairness, it's probably reasonable for him to expect readers to understand "partial credit scaling models," "common item equating," "post-secondary cohorts," and "an item spiralling design" which results in a "sparse data matrix." Afterall, readers are supposed to know <u>something</u>! On the other hand, his allusions to Academic Profiles, "the proposed NALS literacy forms for 1992," and "the GRE data base assembled for NSF," are troubling, especially since his knowledge of them (and I don't mean the acronyms) informs both his conclusions and his recommendations. They sure play havoc with the notion of having a conversation, or forming a consensus.

II.

In the first paragraph of the essay, on his way to

addressing "Issue One" of the guidelines, Dr. Rock says, "Issue one implies a closer tie than probably exists between the skills that are required in the workplace and those that are taught in college." Possibly. Certainly in many academic disciplines this is true. Yet it's an easy thing to underestimate. The SCANS Report is only the most recent voice in a growing chorus. Fortune magazine generally, and the Spring 1990 "Saving Our Schools" issue in particular;[1] the U. S. Department of Labor's Workplace Basics: The Skills Employers Want;[2] Motorola's The Crisis in American Education;[3] Rockwell's Emphasize Education. It's Our Future;[4] Workforce 2000: Work and Workers for the Twenty-first Century;[5] and a variety of other publications have all made quite clear just what "the workplace" wants from education. The extraordinary and growing interest which college and university faculty in general, and vocational and technical college faculty in particular, have shown in critical thinking indicates that at least they are paying close attention. And the University of Notre Dame's College of Business Administration's June, 1991 conference -- "Critical Thinking, Interactive Learning, and Technology," sponsored by Arthur Andersen and Co. and co-planned by Tom Frecka, Chairman of the Department of Accountancy, and Jean Wyer of the prestigous accounting firm Coopers and Lybrand -- indicates that the workplace/college "tie" may be tightening elsewhere as well.

III.

As I read "Issue One," it asks that specific critical thinking, communication, and problem solving abilities be identified with an eye towards achieving "a common understanding and perhaps agreement." A valuable goal. As I read his essay, Dr. Rock does not address this request. In its early pages, the essay seems to imply that some of the abilities to be assessed are, at least to some degree, those imbedded in the already existing NALS, NELS, and NAEP tests, although it also signals that there may not be much there. Later it seems to suggest something of the same for Academic Profiles and, possibly, for the GRE and SAT tests. Elsewhere in the piece, Dr. Rock also talks about having men and women from industry, teachers, and test specialists all work together to define the "KSA's (knowledges, skills, and abilities) that are necessary to perform successfully at each of the [also to be established] performance levels in each of the content areas."

I may be misreading, but I believe Dr. Rock and I just might agree on what the abilties are that should be assessed. As I understand Academic Profiles (to choose just one of the already existing standardized tests), it defines critical thinking as the ability to interpret, analyze, and evaluate material from the humanities, the social sciences, and the natural sciences. Implicitly, the designers of that test signal that critical thinking differs from discipline to discipline; they imply that there is value in defining critical thinking as the thinking done

by the practitioners of the different disciplines. If so, and if that is also what Dr. Rock is signaling by the role he suggests for "professional judges from industry" and teachers in designing the KSA's, then, yes, he and I are in agreement, and I am pleased.

On the other hand, I am not so pleased with Dr. Rock's conclusion that "the proposed assessment of the skills of college students should be carried out with a mix of multiple choice and structured free response. The structured free response being used only for those areas that clearly are inappropriate for multiple choice or other objective methods." Here he and I have a fundamental disagreement. Because I believe the writing process offers the optimal method of assessing thinking, I would reverse the order and opt for either extended free response items, or structured free response items, or, hierarchical structured free response items. At the very least.

As I've indicated in my review of Ed White's essay, my own preference is for portfolio assessment. I define critical thinking as doing the intellectual work of the disciplines, which I believe is best assessed by requiring students to do a variety of discipline-related, primarily writing activities, and the portfolio provides the best opportunity to gather and assess them.

IV.

At one point in the essay, while discussing goal setting in

college level assessment, Dr. Rock says:

> I think that while the consumers of the educational product (industry and government) are reasonably familiar with what skills are required to do the jobs in the labor market, they may not be as aware as they should be concerning the less than perfect relationship between performance on a given item and knowledge of a given desirable skill. (p. 15)

He goes on to say:

> Professional judges from industry, being less experienced with the foibles of testing, are likely to overlook this tenuous relationship between item performance and the knowledge of the skill in question. As a result they are likely to overestimate the performance of students at a given performance level. (p. 16)

Perhaps the limitations are not with the "judges from industry" but with the test? Perhaps the multiple choice test isn't the best way to measure the student's performance. At some point, I'd like to explore this at some length, but, for now, let briefly offer an illustration of what I'm thinking.

At my institution, a community college with a substantial vocational/technical curriculum, we have approached critical thinking across the curriculum by having faculty members, as we say, "unpack" the critical thinking of their disciplines and redesign their courses so that students learn to do that thinking. One member of the faculty, the Director of the Medical Records Technology program, has been doing just that. She met with the heads of medical records departments in some of the hospitals her program sends graduates to, and, together, she and they identified at least some of the thinking abilities required on the job. She has begun to fold teaching those abilities into

the courses in her program. Obviously she needs to assess students' performance of them.

Among the thinking abilities she has identified is the ability to evaluate data and make inferences from it. I've attached (see Appendix A) one activity she has designed to assess how well students can do them. I'd like to suggest that there is little or no "tenuous relationship between item performance and the knowledge of the skill in question." I'd also like to suggest that "professional judges" from a hospital medical records office are not "likely to overestimate the performance" of students who successfully do the task. Needless to say, it's not a multiple choice test.

Appendix A

MRT 130

Health Statistics and Registries

Project: Analyzing Data and Making Inferences

Purpose: To develop skills in analyzing data to make possible inferences; interpreting data.

Evaluative Criteria: This activity is worth 20 points based on completing the steps listed below.

Steps:

1. Review the data on the attached sheet.
2. Develop a list of trends apparent from your review.
3. Analyze what factors could explain the trends. (Be as creative as possible.)
4. Suggest how you would go about determining which factors in #3 would be the real causes of the trend.
5. Submitt a written summary of Steps 1-4. Your summary should be a maximum of 1 typewritten double spaced page.

Endnotes

1. Perhaps the single best introduction to the growing role of "the workplace" in education is this Spring 1990 special issue of Fortune. Back issues are still available: Time & Life Building, Rockefeller Center, New York, NY 10020-1393.

2. Anthony P. Carnevale, Leila J. Gainer, and Ann S. Meltzer, Workplace Basics: The Skills Employers Want, published jointly by The American Society for Training and Development (ASTD) and the U. S. Department of Labor, 1988. Copies are available from ASTD, 1630 Duke Street, Box 1443, Alexandria, VA 22313.

3. The Crisis in American Education is available from Edward Bates, Director of Education--External Systems, Motorola Inc., 1303 E. Algonquin Road, Schaumberg, IL 60196.

4. Emphasize Education. It's Our Future is available from Rockwell International, P.O. Box 905, El Segundo, CA 90245-0905.

5. William B. Johnston and Arnold H. Packer, Workforce 2000 Work and Workers for the Twenty-first Century, The Hudson Institute, 1987.

Review of Development of a Process to Assess Higher Order Thinking Skills

by

Donald A. Rock

Review by Joan L. Herman UCLA/CRESST

While Dr. Rock's paper does not deal with the issue of what specifically we should be assessing for goal five, beyond the general categories of problem-solving, critical thinking, and communication skills, it does offer a number of practical suggestions for how and what types of items ought to be developed. His suggestions are grounded in data-based experience and expected future directions in the National Assessment of Educational Progress (NAEP), National Education Longitudinal Study (NELS), and National Adult Literacy Study (NALS).

Dr. Rock details a number of the difficulties of providing assessment information which is truly diagnostic, although he indicates that NAEP is moving in that direction by providing more detailed information about the nature of student responses and by providing information about the gap between what students know and what they should know. Although I certainly agree that large scale assessment in the past has not paid sufficient attention to the utility and diagnostic value of the information it provides, I think it is a mistake to think that data from a necessarily broad national assessment can be sufficiently detailed to provide the specific

diagnostic information which Dr. Rock envisions. While he believes that assessment tasks must "be sufficiently detailed that appropriate teaching behaviors immediately follow from the definition, I believe that such a level of detail is likely to be counter-productive, particularly given the diversity of offerings in higher education and the strong traditions of academic freedom. By providing information about what students can and cannot do, tests do provide formative and generally diagnostic information. We can't expect a test to tell educators what to do; rather we should rely on their professional expertise and professional judgments to figure out how to solve the identified problems; we also need to think about appropriate incentive systems which will encourage them to do so. Although I disagree with the apparent specificity of Dr. Rock's design ideas, however, I strongly agree with his implicit message that any assessment hoping to provide information about students' proficiency and hoping to provide coherent diagnostic information needs a strong, apriori design scheme (not one defined defacto after test administration).

Dr. Rock's paper describes the types of items and types of scoring schemes that are being developed to address the kinds of higher level thinking skills that will be required for any assessment of higher education. Extended free response and hierarchical structured free response items seems to hold promise, although it may be that even these are insufficient to tap directly students' ability to define, structure, and solve ill-structured, complex tasks. For example, might tasks which are completed over several days, a semester or a year be among the types we should consider? I

heartily endorse Dr. Rock's recommendation that the assessment include a variety of item types; the types included, however, may also need to consider such things as projects, portfolios, etc. I worry that many extended response tasks of the NAEP variety are very time bound and discrete; I also worry that their "diagnostic structure" needs to be thoroughly validated. For example, a study several years ago at CRESST (Webb et al, 198) indicated that students were not consistent in their wrong answer choices; such consistency is a necessary precursor to validating diagnostic patterns.

Like Resnick and Ratcliff, Rock also highlights the problem of motivating students performance on any kind of national assessment of higher education. While he suggests that the choice of item types may influence the extent of the problem -- that is students are more likely to omit free response items -- I think such a solution would have marginal benefits at most. Whether students are not paying very good attention to multiple choice items or explicitly skipping free response items, the problem remains the same: we are not getting a good estimate of their skills and abilities. Serious thinking needs to be done about incentives and appropriate context.

I would agree with Dr. Rock that available measures are insufficient to the task of assessing goal 5. As he points out, the NSF SAT-GRE data base may be the best of what's available, but it suffers from serious validity problems. The Academic Profile data base seems more on the order of what needs to be done -- although on a more representative national sample, and including more complex tasks, assuming the motivation problems can be alleviated

NALS may provide an interim solution -- but in the long run one would hope that there would be serious ceiling effects. (In the short run, these ceiling effects may be less than one would hope). Harris poll types solutions targeting seniors and/or recent graduates and their employers also should be considered. The short term solution would provide time for a longer-term design effort to build appropriate measures and designs, as Dr. Rock recommends. The challenge of reaching consensus on what to assess and of dealing with the motivation issues cannot be underestimated.

GRADUATE SCHOOL OF PSYCHOLOGY
935 EAST MEADOW DRIVE, PALO ALTO, CALIFORNIA, 94303
(415) 494-7477

Evaluation Institute

Michael Scriven, Director

2. DONALD ROCK ON ASSESSING HIGHER ORDER SKILLS

Rock draws on his extensive experience—and that of his institution, ETS—to give us a most valuable overview. He immediately identifies the Achilles heel of the Issue One guidelines, the requirement that the goals be specified "from a teaching/learning perspective". If this is interpreted as he, plausibly enough, interprets it—to mean that "appropriate teaching behaviors immediately follow from the definition", we will have to reject it. We don't even know how to specify the goal of grammatical speech or writing in these terms. But Rock underestimates the difficulty. He implies that this requirement is equivalent to "providing diagnostic information that can inform instruction". That's not impossible, and if that's all the requirement implies, we can do it fairly easily. The impossible part is connecting the diagnosis with the pedagogy. One must understand that the gap between diagnosis and remediation is a *logical* gap, one which cannot always be bridged by empirical discovery; in the limiting case, when the physician diagnoses terminal cancer, one cannot fault the diagnosis on the grounds that s/he hasn't tied it to therapy[1]. There is no successful therapy for many conditions, in CT as in carcinoma. Diagnosis that is *helpful* for the teacher we can reasonably expect; diagnosis connected to *certifiably successful* teaching we cannot expect, apart from a few special cases.

Rock notes that much of the scoring of extended free response items is in terms of right and wrong. I think he is forgetting the vast essay scoring efforts in various states, but to the extent he's right, we should try to avoid the casual recommendation of moving to detailed

---

[1] Detailed support for this view is provided in *Evaluation Thesaurus , 4th edition*, (Sage, 1991), in the entries on Diagnosis, Remediation, etc.

scoring keys. This is jumping into the fire because of the huge cost of developing them, the impossibility of individual instructors developing them with any assurance of validity, the poor sampling of the CT domain they involve, and the horrendous cost of scoring them. His 'super items' represent a step towards improving the yield from multiple-choice items, but are still based on them will all their disadvantages. And, in the CT area, constructing super items would involve some very dubious assumptions about cognitive hierarchies. It is better to move to multiple-rating items (see Appendix, "Multiple-Rating Items").

Rock rightly emphasizes the problem of motivation, in several of its aspects from attendance to test completion. Here again multiple-rating items have an advantage over extended free response items although probably not over multiple-choice items whose faults lie in different directions. But the main brunt of motivation should be addressed by connecting the tests to instruction that produces gains in goals seen as valuable. (Morante's reminders about what did and did not work in NJ are also essential reading on this point.)

On Issue Two, setting standards, Rock says that "the goals must be realistic" in order to avoid frustration by students and teachers. It's clear from his later comments on the NAGB effort in math that by goals he means the standards, and by 'realistic' he means within the attainment range of most or all students, I would disagree strongly. For me, 'realistic' means only one thing; good enough to cope with reality. It is corruption of feedback to students to lower the standards so as to avoid their frustration. Since we'll have a scale, we can compromise by setting intermediate goals for a given instructional intervention; but we should no more compromise on these standards than on the goals in the NAEP math standards setting effort. If it takes mastering this and this skill to cut college first year courses, or do your income taxes on the short form, then that's what the standards should say defines the cutting score for the line between the C and the D grade for the 12th grader; and if what they say is that 74% of the high school population lacks these abilities, then we have to start looking harder at what's happening in high schools (and earlier). The NJ GIS tests will give us plenty of support. There were indeed some serious errors in the NAGB effort at setting standards for math in grades 4, 8, and 12, speaking as one of the external

evaluators that they hired to oversee the project. However, these did not include excessive reduction of standards when the judges heard the figures on the proportion of students that would fail to meet the standards.

On Issue Three, the use of existing data, Rock makes a rather objective claim about the utility of the NALS and ETS data, to which I would only add that it also suffers from weak conceptualization of CT skills. Until this problem is solved, 'developing more sophisticated scoring keys' for the open ended items is going to involve a great deal of wasted time, since better items and better keys are essential. The same applies to any use of computer assisted adaptive tests, but this is not to deny the importance of incorporating this methodology into any large-scale testing effort. Rock is wrong to say that we would be restricted to multiple-choice items for this; multiple-rating items would be just as useable, and considerably better.