

THE *STARNET*[™] MODEL FOR PERFORMANCE ASSESSMENT

by
R. Robert Rentz
Charlotte C. Rentz
R&R Research

StarNET is a vision of 21st century assessment, using 21st century technology, in the pursuit of excellence in human performance. Let us imagine for a moment...

Les Reed teaches auto mechanics to tenth graders at Cedar Valley High School. At CVHS all teachers are responsible for including writing assignments in their classes. Here's a part of the dialog from a recent class.

Reed: "...so class, now that you have finished the final drafts of your articles for the shop newsletter, take one last look at them before you turn them in to me. Remember, these will count on your grade. I'll have your scores for class tomorrow."

As the students turn in their papers and leave class, Reed walks to the corner of the classroom and removes a videotape from one of the permanently mounted camcorders. This is the camera that is synchronized with the infrared sensor on Reed's tie-clip microphone. The sensor permits the camera to focus on Reed as he moves around the classroom. The other camera, in the front corner, has a wide angle focus and captures phenomena like student reactions and aspects of the culture of the classroom.

Les Reed takes the videotape and the student's papers and starts down the hall to the assessment center. On his way he stops by Peggy Murray's classroom. Peggy teaches Spanish.

"Hi," says Peggy. "I appreciate your processing Alan's audio tape for me; I just have to work on these assignments. I'm real pleased with the speech Alan gave in class today, he got several laughs,... all in Spanish," laughs Peggy.

"No problem, but you owe me one."

Les takes Peggy's tape of Alan's Spanish speech, walks into the assessment center and sits in a little booth, surrounded by computer peripherals. Les enters his ID and password using the bar-code wand to scan his ID card. He inserts his videotape into the video unit, queues up the starting point of the lesson he wants evaluated, clicks his mouse on the video icon, and sets the stop time for 10 minutes.

While the video unit is transferring from tape to disk, Les clicks on the document icon, inserts his class's papers, clicks on go, and waits the 30 seconds for the 25 papers to scan. The scanner processes 40 pages a minute, and the video unit runs at high speed,

so he's finished in two minutes, but he has to enter some additional instructions for the StarNET system, and that takes a few more minutes. Les then inserts Peggy's audio tape into the audio unit, scans her ID card, and processes the cassette.

He's been there less than 10 minutes, and as he leaves, he knows the students' writing results and a report evaluating his videotape lesson will be in his electronic mail when he gets to school the next morning. Peggy's reports will be in her e-mail as well. Now they both can get to work right away on instructional activities that will increase their students' performance, and Les will also have information to assist him to improve his own performance in teaching.

Performance assessment is an evaluation activity that delivers its greatest service in support of instructional and intervention efforts. Good instruction and intervention results from a sharp focus on performance and timely feedback about what individuals know and can do. Getting immediate and correct information about what individuals know and how well they are performing helps decision-makers decide, bosses manage, and teachers teach. It is the key to improved human performance.

Performance improvement requires that we can define desirable or expected performance. Then upon observing actual performance, we can compare it to our expectations. This observation and evaluation of performance is what we call assessment. That assessment data then is used to tailor intervention to improve performance in those areas previously identified and assessed. Of course, the process requires another assessment then to determine if performance has changed since the previous evaluation. Thus the assessment-intervention cycle is repeated on an ongoing basis to track performance improvement or to monitor levels of performance.

The StarNET vision places a high premium on implementing seamless, authentic performance assessment. To do so, the assessment exercises and associated responses should replicate as closely as possible that which would be observed in "live" action assessment. For example, assessing a foreign national's oral English skills is best done by actually evaluating samples of that individual's oral presentations and the individual's responses to oral requests from others (perhaps on audio or video tape). Such assessment would be much closer to live action assessment than would be the case, if the individual were answering questions on a paper and pencil test in which 1 of 4 possible answers is selected for each question.

Authentic assessment then demands a multimedia capability, i.e., the capability for capturing performance information via written text, photographs, audio recordings, and video clips. Seamless authentic assessment also demands that evaluators, instructors, and supervisors have a shared understanding of different performance levels. Furthermore, seamless assessment requires frequent or periodic assessments, conducted in non-intrusive ways, and resulting in timely data for decision making. High quality, rapid scoring and reporting becomes critical to successful seamless, authentic assessment.

Performance assessment has traditionally been viewed as very expensive because of the use of raters in scoring rather than the use of machines for scoring. In performance assessments, one prevailing characteristic is that reliance is placed on human observers making the judgments, and if the performance assessment program is high volume or large scale then it is not only expensive, it is typically cumbersome and slow, with results getting to decision makers well after the time when any productive use could be made of those results.

StarNET, a computerized system for managing performance assessment, combines assessment methodology (particularly training and monitoring of raters) with computer digitizing and networking technology. StarNET can not only deal with frequent and high volume multimedia performance observations, it can also provide inexpensive, rapid turnaround of the evaluations of those multimedia observations in order to effectively influence decision making. The StarNET methodology of integrating assessment and technology results in a system which performs three major functions: 1) capturing and digitizing examinee's performance on assessment exercises at the point of origin, 2) assigning resulting digitized data to appropriate evaluators for evaluation, and 3) returning that evaluation to the source.

For a given client, *StarNET*[™] is the totality of the hardware, software, assessment, and human systems that work together to produce authentic and seamless performance assessment in that client's situation. Individually customized for each client, all StarNET systems are implemented with state of the art, industry-standard methodology and technology, and all systems are designed both to maximize the use of existing client and industry resources and to ensure ease of future system modifications or upgrades.

StarNET integrates assessment and technology into a system for performance improvement that not only makes seamless, authentic assessment possible, it results in a program that is financially viable, operationally feasible, and timely for high volume ongoing programs. Four components comprise a StarNET application:

- Local Multimedia Data Collection Center(s) (e.g., hardware configuration, data capture, digitizing, & transmission software)
- Management Hub (e.g., management & communication system hardware & software, hub location, rating & reporting function procedures)
- Rater/Scoring Interface Unit(s) (e.g., multimedia-based computerized scoring, scoring interface systems, rater unit workstations, rater training & monitoring)
- Communications Network (e.g., network configuration, distribution system, hardware & software)

Since each StarNET application is intended to maximize the effectiveness of assessments in fostering the success of a local performance improvement program, the design for the StarNET system must be customized to each individual client's resources, priorities, and constraints. As such, local design factors require identification and incorporation into each StarNET application. In particular, four areas are explored in designing each StarNET system:

- Existing Assessment Program (e.g., exercise bank, statements of standards, renewal & quality control procedures)
- Scoring and Monitoring Considerations (e.g., scoring system, training approach(s), quality control procedures)
- Uses of Assessment Data (e.g., staff resources, administrative priorities, prior use of assessment data in determining intervention strategies)
- Infrastructure for Implementation and Sustainability (e.g., current technology and planned hardware and software acquisitions, budget allocations for assessment, professional development, and technology)

StarNET specifications review options for each of the four system components and customize the system to best meet a particular client's constraints and goals. Those specifications might be as simple

as a flow chart integrating existing capabilities and suggesting staff responsibilities and time frames. On the other hand, StarNET specifications for a client could be a much more elaborate detailing of assessment development, hardware acquisitions, software modifications, and staff training.

Performance Improvement Technology

The StarNET system for managing multimedia performance assessments uses technology to implement authentic assessment, seamlessly by electronically transmitting and receiving digitized performance assessments on both the information super-highway and on other specialized local communication networks. The benefits can be summarized as follows:

- Fast turnaround of scoring results,
- Eliminating the need to move massive amounts of paper from place-to-place for scoring,
- Eliminating the need to pay travel expenses for raters,
- Permanent, inexpensive storage of performance assessments & portfolios, and
- Central office auditing of assessment results.

As our 21st century assessment illustrates, StarNET is applicable to those classes of examinee responses that can be digitized and transmitted electronically. The number of different types of performance assessment responses that can be digitized and represented on a computer monitor is quite large and that number increases every day. Thus, the number of performance responses that are appropriate data for StarNET is also large.

An individual's performance on an assessment exercise is captured and digitized at a local multimedia data collection center and that digital object is then transmitted across the communications network to a trained evaluator, and the evaluation is then returned to the source using the same communications network. The entire process is controlled and monitored by the central hub management and communication system. They are described in more detail in the following sections.

Local Multimedia Data Collection Centers

Capturing and digitizing examinee's multimedia performance on assessment exercises at the point of origin is the function of Local Multimedia Data Collection (LMDC). An LMDC center consists of one or more booths or similar workstation areas surrounded by computer peripherals. The booth area might occupy floor space of about 5' by 5' and be about 5' high. You could be sitting at something similar, even as you read this; for example, you may be sitting at a "U" shaped work surface, with a monitor and keyboard in front of you. To your left you might have a document scanner, to your right a laser printer. You may not have a video or audio transfer unit connected to your system, but you could. All the technology exists today to establish Local Multimedia Data Collection Centers that will take advantage of all of the StarNET capabilities.

The major function of a Local Multimedia Data Collection Center is simple: it provides for inputting assessment data, transforming that assessment data into digital form, transmitting the data to the next component of StarNET, and receiving the resulting reports. Each workstation contains a computer and the peripherals necessary to transform assessment data into digital form. The Local Multimedia Data Collection Center would start with a PC or desktop workstation. Peripherals would include a document scanner for scanning single pages of paper, such as a student's essay or a file of pages from a portfolio. Document imaging systems that scan 100s of pages in a few minutes have been available for some time

and each year they just get better, faster, smaller, and cheaper. The hardware found in a full function multimedia data collection center would include the following:

Computer	Modem	VCR
Keyboard & Mouse	Scanner	Video Edit Monitor
Printer	CD-ROM (Read/Write)	Audio Cassette

Together, the modem and the scanner comprise an inexpensive desktop, document imaging system which allows realizing many of the advantages of the StarNET system without the necessity of implementing the full blown system. A digital color camera for taking still photographs can be added to the LMDC center. Cameras whose output goes to a floppy disk can be transferred directly to Star NET. The Local Multimedia Data Collection Center incorporates players for both audio and video tape that are connected to the PC through expansion boards that capture, digitize and compress these assessment performances. Components can be added or replaced as the technology advances. The quantity of data that can be handled in this arrangement is potentially unlimited: the types of data objects about which the StarNET system is concerned require megabytes of storage.

The StarNET software for the LMDC center is straightforward in purpose: to maximize the effective functioning of the multimedia data collection center tasks of data capturing, transmission, receipt, and storage. For the most part, software supplied by hardware vendors will suffice. Besides an appropriate user interface, the hardware will determine what software is necessary. The technical problem here is integration, choosing components that work together. Choices exist today that accomplish the functions discussed here.

As the CVHS example illustrates, StarNET is applicable to those classes of examinee responses that can be digitized and transmitted electronically. Examples of such responses include: scanned images of a student's essay; video tape of a dancer's performance or of an instructor's demonstration of carpentry competencies; an audio record of a poetry reading or a public speech. Other examples include photographs of a science project or a computer program written directly on another computer and transferred via floppy disk to the Local Multimedia Data Collection center. In other words, any performance assessment response that can be digitized and represented on a computer monitor is appropriate data for StarNET. Current technology supports digitizing documents, photographs, audio, motion video, animation, not to mention ordinary ASCII text, and computer graphics.

The Local Multimedia Data Collection center is very flexible and adaptable. It can serve both nonassessment purposes, such as administrative and instructional functions, as well as assessment purposes. An additional feature of the Local Multimedia Data Collection center is its potential mobility. A workstation can be installed in a van and driven to several sites that might share it. The degree of mobility will depend on options that are chosen for the communications network.

The StarNET LMDC center design for a particular client includes detailed specifications for hardware and software to accomplish the client's goals in performance assessment areas, and includes a review of uses for LMDC center resources in high priority nonassessment areas. The recommendations included in those specifications take into consideration existing hardware and software resources as well as fiscal and other local constraints.

Management Hub

In the StarNET system, management and the associated responsibilities for scheduling, monitoring, and reporting are assigned to a Central Hub. The overall role of the Central Hub is management. At the Local Multimedia Data Collection Center, the StarNET system creates multimedia database records, tagged with the necessary codes, and the system then begins transmission of the database records to the Central Hub. The Central Hub operates much like an on-line transaction processing system, receiving the multimedia database records, identifying the types of records, storing them for archiving, and routing these records to the Remote Rater Units for scoring.

The Central Hub software identifies which Remote Rater Unit will be assigned to rate each multimedia record. The identification process takes into account schedule, workload, rater calibration (using multi-faceted Rasch models - important to ensure scoring equivalence) and any other factors required for the particular assessment response being scored.

It then receives the resulting ratings and incorporates them into a report which is sent back to the sending LMDC center. Storage is a second function of the Central Hub. Whether this storage is temporary or permanent is optional, however, as was pointed out above, StarNET produces vast amounts of digital data. Another function of the Central Hub is to manage rater training and monitoring, subjects which are covered elsewhere: see the discussion of Remote Rating Unit which follows.

Excluding the hardware necessary for a specific communication mode (e.g., a satellite dish and transmitter, an ISDN line, etc.), the hardware components for the Central Hub consist mainly of a computer, a communications server, and mass storage devices like an optical jukebox. The staff necessary to manage such a system is small and the physical location arbitrary, dependent only on the Communications Network.

Three software subsystems support the Central Hub's functions. The first is a multimedia database management system, of which, there are numerous possibilities with existing off-the-shelf products. The rater routing, management and training subsystem is necessarily customizable to particular assessment requirements, but StarNET provides a shell structure, outlining the content specifications an assessment should include. The third subsystem involves analysis and reporting. These products should also be customized for particular assessments, but StarNET does provide a library of reporting options a client could select from.

Central management functions can be designed for a Local Management Hub configuration or can be assigned to the StarNET Central Hub. In some cases, a client may wish to retain some management functions locally and allocate others to StarNET Central. This is one of several scenarios for managing rater training and monitoring when a Local Management Hub model is chosen by the client.

Rater/Scoring Interface Unit

In performance assessment the sole function of the scoring system is to judge examinee performance, that is, to rate the individual's performance on assessment exercises. If the scoring is multimedia and computerized, a Rater Unit consists of a trained human observer, sitting before a multimedia PC, connected to a Central Hub through a high speed data link (probably a telephone line). The physical

location of the Rater Unit is arbitrary, it can be anywhere in the world (e.g., local or remote to the location of the of the examinees). With standardized training and monitoring utilizing interactive multimedia and StarNET system components, there is no necessity for raters to gather into groups, no necessity to provide workspace or pay travel expenses. A talented workforce can be assembled without regard to geographical boundaries, time zones or a nine-to-five schedule. Quality ratings can be delivered like business inventories, "just-in-time."

Scoring Interface System

Once raters are qualified and become proficient in the use of the scoring interface systems, then actual scoring can begin. Student responses are scanned and the images written to CD-ROM's. The examinee responses and rater scores are delivered and collected, respectively, using computers. Either a Remote Rater Model or a Local Scoring Center Model may be used in establishing the scoring interface system.

The Remote Rater Model involves rating units established anywhere highly qualified raters are located; e.g., individually or in small groups at home, offices, or other locations in whatever state or local they reside or work. Standardized training and monitoring via the StarNET Communication Network means that space and travel costs are not incurred.

The typical Local Scoring Center is an Intranet Scoring Center and is usually comprised of nine to 12 workstations. A database of student reference numbers is used to randomly deliver exercise responses to raters. Further, raters are not assigned a work station, so the station at which one sits varies during the project.

The scoring center operates days and evenings. This schedule has been shown to be beneficial to raters and clients. The number of raters for any shift is limited to the number of workstations. It is important to the client to keep the scoring center full of raters who are ready and willing to work. This flexible scheduling serves both purposes.

Once raters are scoring, daily rater statistics are calculated using a 10% random sample of responses, reassigned (randomly) to a rater other than the original rater. Because of these second readings, any raters whose scores vary from the rest of the group are easily and quickly identified and retrained.

This computerized scoring system is entirely run with Internet compatible programs and procedures. The StarNET experience in past applications is that although some raters may be initially frustrated by indirect manipulation of electronic "student papers" and the computer input, it is no time at all until no rater would return to a paper and pencil scoring method. With ongoing updates to the system the process increasingly becomes more and more "user friendly."

Computerized scoring has several advantages over other techniques. Data entry is "real time." Because raters enter their scores, the database is updated as soon as the rating is made. The database is always up to date. This procedure minimizes errors because it is direct; the rater selects the score(s) for each response and submits it to the database.

Another advantage of this image-based scoring interface system is that it is designed to prevent human error during scoring. When raters are assigning scores, several built-in features will prevent the simple errors which can occur in paper-based scoring. For example, to assign a score, the raters will simply

use a mouse to select a "button" on his or her screen. Only scores which are valid for that item will appear on the screen. Thus, it is impossible for a reader to assign a score which is outside the valid range. It is also impossible for a reader to pass over an item without assigning it a score, as the image system will not distribute more work to a reader until the item currently on screen has been scored.

This system also reduces time to score. Because there are no papers to shuffle, no packets to physically pick up and put back in some central location, time is saved.

Rater Training and Monitoring

Perhaps the single most important aspect of producing quality information in a performance assessment setting is the training of raters and the monitoring of their subsequent ratings. Raters must either maintain the same standard for the ratings they assign, or, if they change, that change must be detectable and correctable. Discussions in the writing assessment literature about rater drift, retraining, rescoring and equating stem from this source of error. If comparable rating results, from one assessment to another, are to be achieved, whether over time or over examinees, training must be standardized. Furthermore, we need to be able to detect rater drift as soon as it begins. StarNET provides for standardized training and rater adjustment.

Training options for the client allow training using CD-ROM or training in a traditional workshop setting. The major advantage associated with the traditional workshop approach is that participants view the involvement and interaction with other participants as valuable. Expense and lack of substantial individualization of training are major disadvantages of this approach. Presented as an interactive multimedia program on CD-ROM, such training would be individualized, effective, and relatively inexpensive. The workshop approach, however, is frequently chosen when the client adopts a Local Intranet approach to actual scoring.

One advantage of using Remote Rater Units is that they require rethinking the concept of rater training. If raters are located anywhere, they must be trained anywhere, and that realization forces an individualized approach to training. What raters need to do is examine a particular assessment and to assign the correct rating. To do that they have to practice that specific task. They must understand the task and then practice on a sufficient number of pre-rated examples (whose correct rating is known) to become proficient. Interactive multimedia lends itself well to this requirement. This method uses the same equipment that raters use for rating. Using a training-by-example model or some other acceptable training approach, training content is produced, and published on CD-ROM, along with pre-rated carefully chosen examples.

One outcome of training is rater calibration. Studies are beginning to appear pointing out the advantage of using multifaceted Rasch models to calibrate rater banks along the lines of item banks. StarNET is a natural vehicle to use these methods. Rater calibrations are used to select the subset of raters to rate a particular assessment response, as well as, to make final adjustment to ratings for the purpose of achieving equated scores. Rater fit statistics provide useful information about the need for retraining and that process begins as it is detected.

The content of a training program is specific to a particular assessment and customized to local standards. The added benefits of creating an individualized, computer based, training program include the potential for use in a staff development for instructional and supervisory staff; and the necessary equipment would already exist at the Local Multimedia Data Collection center.

Rater performance monitoring is automatically incorporated into the management functions of the StarNET Central Hub. When a Local Management Hub model is chosen by the client, monitoring can be designed into the Local Hub or, as is most frequently the case, rating monitoring responsibilities are contracted out to StarNET Central. Whatever model is used, monitoring procedures involve embedding pre-scored or double score performance responses in the scoring process.

Only two hardware components are needed by the Remote Rater Unit: a PC and a telephone line. The PC should be equipped with a CD-ROM reader, color monitor and high speed modem. Assessment responses might be delivered in batch (stored on CD-ROM) or transmitted via modem. It is interesting to note that while the size (in bits) of the digitized assessment data is large, that data only need to be transmitted one-way. Only the ratings themselves have to be returned.

The Remote Rater Unit requires three software subsystems. The first is a database management system that allows viewing of the assessment object, provides for data entry for the resulting ratings, and transmits those ratings back to the Central Hub. Another subsystem is one that accommodates training programs that will probably be distributed on CD-ROM, with their own specific software. The final subsystem is management related and allows the rater to provide information (e.g., scheduling) to the Central Hub.

Communications Network

The function of the Communications Network is to move digitized assessment data (i.e., multimedia database records) between StarNET components. The requirements include handling massive amounts of data and high speed. The central idea of StarNET is to move digital representations of assessment responses to well trained, human observers, to report their evaluations, and to do this task fast.

The Communications Network component of StarNET is conceptualized a bit loosely. Specific hardware and software components of the Communications Network undergo rapid development, and what is available one day can be outdated six months later. Wide area network services proliferate. Applications like video conferencing and EDI systems demand larger capacity and faster communications services. In short, the telecommunications field changes constantly and will be different at the time a particular client adopts the StarNET system.

In some situations, the StarNET Communications Network is best defined as a local area network (LAN), with local staff operating the Rater Units. If the Communications Network is a LAN, then the Central Hub could exist only in software (a virtual Central Hub!). A computer that is a component in the LMDC center can perform the functions of the Central Hub. The customized design, in such a case, might conceptualize the LMDC center as just another server on an existing LAN, and almost any LAN compatible workstations could be configured to work as RU's. Combinations of existing and new equipment can be used to minimize cost. If the Central Hub were a LAN, than a variety of existing hardware can be incorporated into the system design.

Early implementations of StarNET have used two kinds of distribution between the Central Hub and the Remote Rater Units. One is batch, where a number of assessment records are stored on CD-ROM (6000 pages of student essays per 5.25" platter) and delivered overnight to the Remote Rater Unit. The other method transmits a single record at a time to the Remote Rater Unit over the Internet.

In the case where volume of data is greater, the communications like between the Local Multimedia Data Collection center and the Central Hub requires a different strategy. The technology includes satellites, etc.

Whatever Communications Network is configured for a client, the design taken careful consideration of the amounts and kinds of data to be transmitted, the existing hardware, software, and local networks in use by the client already, as well as other non-assessment uses of the communication network (e.g., e-mail).

Local Design Considerations

Assessment Program Resources

The beginning point in a StarNET application is the assessment data. Performance can be conceptualized as written, spoken, or visual products and thus best be captured in written text, photographs, or audio or video clips. However defined, tasks which stimulate the required performance, and accompanying standards which define acceptable levels of performance, must be specified. For each performance standard there may be many different tasks which would elicit a performance response which could then be evaluated against that standard. In some applications, it is also the case that more than one set of standards may apply, e.g., national, and local, each having been defined by professionals or policymakers in the field.

Assessment development, analysis, & research plans are derived from an assessment design or framework. It is through such a framework that content and skill specialists define performance indicators: types, levels, and products of performance which qualify as meeting a desired standard. Through the assessment framework, standards are classified, performance indicators are categorized, and content/standards groupings are indicated. What scores are desirable at what level of accuracy (individual or group) is also determined from the framework.

Exercise development & review, tryout of exercises, exercise selection based on tryout data, assessment instrument construction, instrument pilot, equating, scaling, and reporting all flow from the assessment framework specifications. The assessment framework defines the structure & classifications of exercises in the pool, whatever the exercise type: selected response(SR), constructed response(CR), short answer(SA), extended response(ER), performance events(PE), or performance tasks(PT). The number of exercises, the structure of test forms, as well as what instruments are needed and whether the items and forms should be the same for all categories of students are determined by reference to the assessment framework.. All long range plans for managing exercise development & sustainability are guided by the framework.

Initiating exercise development, developing a scoring system, and designing a management system are dependent upon the framework specification; nevertheless, work can begin at the onset by first building a temporary framework using an existing exercise pool to initiate development of performance indicators and standards as well as the development of new forms & assignments for item writing. Whatever the process, the development of the assessment framework is critical.

Scoring and Monitoring Considerations

Scoring System Definition

A major component for any performance assessment is the definition of the scoring system. The scoring system requires a definition of the content and skills which are to be assessed and a rubric is developed which provides the conceptual statement of the quality of response that gives a participant a particular score. Also called a scoring guide, the rubric is a set of guidelines for scoring student work. A typical rubric states the assessment criteria, contains a scale, and helps raters, instructors, and supervisory staff rate student work according to the scale. The basic rules for assigning score points does not require a defensible operational definition of the score, however, anchor responses can become an integral part of the definition of the score points. An illustration of a rubric can be found in Exhibit B.

If the client has already formulated standards, those standards are incorporated in an assessment framework and provide a theoretical notion of the dimensions for scoring performance tasks. In cases where articulated standards do not exist, client selected content and skill experts must be assembled to define the dimensions for scoring and a set of rules for assigning scores to responses to performance exercises. It is also critical that the scoring system be client approved before training and scoring begins.

In previous StarNET applications, an Anchor-Exemplar-Practice Response Model has been used for both the training of staff to rate performance and the training of instructors to use assessment results to improve performance. Such a model involves selection or construction of three types of performance responses: anchor, exemplar, and practice responses. An anchor response is definitional. The anchor performance defines what the score point means. An exemplar response is nearly an anchor response; it is clearly an on-point performance but with some obvious undesirable characteristics. Such responses provide an elaboration of score points and are used in the early stages of training and retraining. After presentation of anchor performance, the next 5-15 performance responses presented for training are exemplars. Practice responses are selected or constructed to represent and span the whole range/distribution of responses/score points likely in the population. When and how they are presented in training sessions is determined based on how far the practice performance deviates from the anchor.

Anchor pulling sessions are held with content and skill specialists during which actual performances are reviewed. The objective is to locate performance responses in each of the anchor, exemplar, and practice categories which can be used in training. When suitable performance responses cannot be located, responses which meet the criteria must be constructed. Such constructed responses must meet the rubric requirements and simulate actual student performance (e.g., if the performance is a writing sample, the paper must look as if it were written by the population being scored).

Local considerations may require some modification to the approach described or a complete redefinition of the scoring system to be used. The particulars of the scoring system are not critical in StarNET. What is critical is that a locally approved scoring system be used in designing the training for raters and that that system contain clearly defined scoring guidelines. StarNET specifications review the nature of existing client resources for scoring and outlines the steps for finalizing the scoring system definition.

Training and Monitoring Raters

StarNET provides an economically viable, productive scoring approach to performance assessment. StarNET staff plan, program, hire raters, and monitor computerized scoring. StarNET trainers have experience training raters to score performance exercises in various content and skill areas and monitoring raters as they apply a rubric across all the varied responses generated in practice.

Hiring. Reliable raters and their training are crucial components to successful measurements of performance exercises. The first step toward this success is hiring. Targeted advertising and thorough screening and interviews are strategies which are used to ensure the maximum number of individuals who are successful in scoring. Raters with previous experience in scoring performance responses and on-the-job-experience are more familiar with, and have a clearer concept of the work.

Rater qualifications include: on-the-job-experience, acceptance of computerized system, appropriate credentials (degree, certification, or on-the-job experience), and successful completion of training. The StarNET approach accommodates raters' needs and requests and thus involves interesting projects; flexible work hours and days; comfortable facilities; professional, personable workplace; support and leadership; good pay; and humor, on occasion.

Whether some or all of the raters are selected from the ranks of the client's staff will depend upon factors such as the client's desires for local control and management of various or all aspects of the StarNET system once implemented. Those requirements are incorporated into the StarNET specifications as parameters within which the recommendations are made.

Training Options. Since performance assessment requires human raters making judgments, rater training is the most important component to ensure quality measurement. Effective training is the second component in successful scoring. The training provides intensive instruction and practice in those content and skill areas defined by the rubric. At the end of training, raters effectively apply the rubric producing highly consistent scores for a given response. Thus, the raters perform their tasks so that measurement is valid and reliable. It has been the StarNET experience that trainers and trainees, working as a team, quickly operationalize the rubric whatever response is encountered.

A number of different scoring and training models are available for use in performance measurement. Prior StarNET applications have most often used a "Training-by-Example" model, and it is that model that is used in this discussion for purposes of illustration. Whether that approach is recommended for a particular client depends on such factors as staff experience with, and attitudes toward, other approaches as well as the training-by example approach. The figure below outlines a generic model of the content of a training-by-example program.

Training-by-Example Program Content

- Multimedia Description of
 - Assessment content
 - Scoring dimensions or categories
 - Scoring scale or rubric
- Annotated Examples (Anchors or Models)
 - Model responses for each score point (dead-solid-perfect)
 - Explanation or rationale (text or voice)
- Practice and Qualifying Responses Presented

- Practice Set 1 - Clear (on-point) examples
- Practice Set 2 - Increasingly ambiguous examples
- Practice Set 3 - Typical distribution of examples
- Calibration Set - presented when appropriate
- Re-Calibration Responses
 - Responses used as "check-sets" to monitor raters

The anchor, model, or example responses are always chosen or constructed to clearly represent and define the score scale, and the annotations explain why these example responses are quintessential representatives of their score point. Such explaining is a function similarly performed by a talented human trainer. Rater training can be delivered either through the conduct of "live training" workshops or by utilizing other technology based approaches; for example, through the use of interactive multimedia CD-ROM program, where annotations might be presented in pop-up-windows, as audio clips, or in other multimedia modes. The value of substituting annotations in a CD-ROM program for interactions between the talented trainer and the trainees lies in the standardization of the information received by all raters. Good teachers of trainers cannot help but customize their instruction to their students, a noble practice for heightening interest but a disaster for quality measurement. Quality measurement demands that *finished* raters be as much alike as possible.

A CD-ROM program description can be combined with the annotated examples and presented as a stand-alone videotape and shown as a general information piece or part of a more fully-developed, staff development program. On the other hand, the client may have other considerations in mind if a preference is voiced for "live" training, especially if the raters are members of the client's staff. Regardless of the media on which it is presented, this description is the common orientation received by all raters and the StarNET specifications will provide training recommendations that meet the client's requirements.

After the score scale is defined with both abstract definition and concrete examples, practice responses are presented to the trainee rater. The trainee examines and rates the response using the data entry protocol to be used for real ratings. The main difference between training and "live" ratings is that following the ratings, the training program informs raters of the correct answers and provides justification for that answer. After a predetermined level of accuracy on the practice responses, a set of calibration or qualifying responses is presented; a specific standard of accuracy is required on the ratings of these responses before a rater is allowed to rate real assessment responses.

Monitoring. Assuming a trainee *passes* the qualifying set of responses and is allowed to rate *real* responses, periodic monitoring of that rater's performance is desirable. In writing assessment, for example, papers presented to raters for this purpose are often called *check-sets*. These papers are pre-scored and function either to detect rater drift or to change the rating behavior of a rater whose drift has been detected by other means (such as degree of agreement with other raters). These re-calibration responses are important ingredients in monitoring rater performance. Whether these recalibration responses are stored with the other training content or presented to raters as if they were real responses, is a matter of specific program design requirements. Nevertheless, it is significant to note that responses used as check-sets, within the digitized StarNET system, have the physical appearance of any other real response. Such is not the case for check-set papers often used in typical writing assessments that are photocopies and clearly distinguishable from real papers.

Uses of Assessment Data

When the system developed to score performance assessments is based on the standards, then these standards provide a powerful goal around which quality instruction can be structured & local curricular can be developed. The integration of standards into the design and scoring system for the instructional and assessment products of a performance assessment program provides the opportunity to focus both instructional and assessment activities on the standards. To assist instructional & supervisory personnel in this focus it is essential for them to understand the standards and to be able to recognize when their students have demonstrated those standards. By being able to identify behaviors using the skills embodied in the standards, instructional personnel can better design and implement instruction that directly addresses these standards.

Being trained as an expert rater of the assessment instruments has been found to be an effective method for learning the concepts of standards and learning to recognize when students display evidence of being able to use the skills embodied in those standards. It is not necessarily reasonable to train instructors or supervisors to high levels of expertise as raters. However, in sessions designed for professional development, it is possible to achieve the benefits that rater training offers by engaging in an abbreviated version of the same training that a potential expert rater might go through. When abbreviated training is combined with brief practice, the result is a concrete understanding of the goals of instruction as represented by the standards. When practice is combined with examples of student performance with which the participants can identify, for example, their own students or employees, then interest is enhanced. Moreover, when that practice involves using technology that might be either novel or perceived to be a valuable skill, then interest and enthusiasm is further enhanced.

The scoring system may be complex and not easily mastered by simply reading about it. If the client is interested in actually applying the scoring system to individual's responses, some amount of guided practice will, more than likely, be necessary. Often choices about what kind and how much training can be delivered depend on the amount of time available. For example, what might one do if one only had 1/2 day for training? What about two days? As noted elsewhere, scoring proficiency probably depends on practice, so proficiency and time are closely related.. The first example shows an agenda for the training of raters, where the criteria are high levels of proficiency. The second agenda shows what might be done in one day and is probably enough to give most staff insight into the scoring system and individual's levels of response. The third example requires 1/2 day and is useful for awareness or when knowledge of the scoring system is necessary for other purposes such as exercise development on the meaning of the client's performance standards. The tables show topics with the approximate amount of time to devote to each.

The training can be conducted in a workshop setting where one segment is devoted to Rater Training and a second segment to Practice Scoring. Although the particular agenda items vary depending on how much time is devoted to each of those two segments, in each case, the initial segment involves "training-by-example" where participants practice on prescored responses until the group gains a minimum level of proficiency in using the scoring system. In the second segment, participants actually score samples of student responses, getting feedback on their rating performance. The responses of students who are similar to those encountered by participants are used as the responses on which the workshop attendees practice.

One day for installation and testing of software is typically added to the time allocated for these training activities. The optimum number of participants for each session would probably be 25. The workshop uses computer workstations, one for each participant. Thus the workshop depends on the

availability of a site where there are at least 25 IBM PC workstations that are networked and use Windows95 operating systems. Detailed specifications are provided when the client wishes to add a component to the system to help staff use assessment data and procedures to improve performance.

In some cases, if the client want to emphasize using assessment results to design intervention strategies, a third session on Instructional Planning has been added. In this third session, the workshop participants apply their newly acquired skills in recognizing student skills and performance quality to design instructional activities focused on the standards.

Some clients wish to add a follow-up workshop which focuses on having participants share their experiences and presenting data from their students. If it is possible to administer a survey to students prior to the workshop, the data from that administration is discussed, with an emphasis on using the data for evaluation and subsequent activities. Some time is devoted to planning for the following year. Often the workshop is planned for 2 to 3 days.

Infrastructure for Sustainability

An infrastructure for the sustainability of an ongoing program to improve performance in any setting includes:

- a bank of assessment tasks matched to agreed standards of desired performance
- an assessment system technology and methodology which allows timely evaluation and reporting of actual performance relative to those standards
- a quality control and renewal system to ensure continuing accuracy of evaluations and
- replenishment of assessment tasks for the bank
- a strategy for using data from the assessment system to determine intervention or instructional content and strategies tailored to improve performance in those areas found deficient .

The StarNET system begins with, or assumes, an existing formal or understood set of standards of desired performance and perhaps an initial set of assessment tasks which may or may not be matched to the standards. StarNET makes provision for defining standards where they are not yet formalized and for constructing or selecting assessment tasks to match those standards. When the situation requires, a full scale assessment development is planned. In many cases, however, the required assessment resources (e.g., standards, assessment tasks, tables cross referencing standards and tasks) are already available and what is required is defining a scoring system or rubric which delineates the scoring dimensions, and identifies the scale(s) to be applied to each task.

An assessment system which allows timely evaluation and reporting of actual performance relative to standards must include a scoring and reporting approach which provides reports back to practitioners and decisions makers as quickly as possible and certainly prior to the deadline set for their decisions to be made. To do this, any assessment system must allow individuals who collect data (those who assign performance tasks and than collect the results) to submit that information for scoring as soon as it is acquired. Further, results must often be returned to them within hours or a few days to be useful in making decisions and intervening to improve performance. In order to utilize different performance tasks and different combinations of performance tasks (assessment forms or instruments) for different individuals and on different occasions it is also necessary to have compiled exercise and instrument statistics and to have scaled the exercises and instruments. Finally, The system must be financially and operationally feasible.

Capturing and returning the data in a convenient and timely manner, as well as the requirements issues of financial and operational feasibility are the cornerstones of the StarNET Performance Improvement Technology. Analyses leading to exercise and instrument statistics and scaling, if needed for a particular client, are the subject of discussion and recommendations included in the StarNET design in response to local client considerations.

If the performance improvement program is to continue beyond a one-shot attack on the problem, StarNET as initially installed, provides the basis of the structure for sustainability. Whether the client decides to choose a local scoring model or the remote rater model, there must be in place, or in the plan, a quality control and renewal system to ensure continuing accuracy of evaluations and replenishment of assessment tasks for the bank. In the remote rater model, StarNET Central monitors the accuracy of the raters and drafts the plan for task replenishment and scaling of new assessment tasks. During the initial assessment development scoring, StarNET offers the option of adding quality control scoring in which training is provided that is aimed at creating the capacity to score responses in the future, e.g., the next year. If the local scoring model is chosen, StarNET offers assessment update workshops which meet quality control and task replenishment requirements. A local sustainability program is also available from StarNET which trains the local staff to maintain the performance improvement program locally and to conduct local update sessions themselves.

In the remote rater system, the scoring quality is monitored using two different models. The validity model uses pre-scored papers which are included in the work assignments routinely distributed to raters. The double-scoring model involves a second reading where a second rater scores a particular task response after that response had been scored by an initial rater, and the results are compared. StarNET uses the data collected in these ways to identify drifts in scoring quality and intervenes with additional scorer training when indicated.

Intervention strategies are the responsibility and prerogative of the client. Upon request from the client, StarNET staff can be made available to assist in developing or implementing a strategy for using data from the assessment system to determine intervention or instructional content and strategies tailored to improve performance in any areas of performance found deficient.

The StarNET design developed for a client always had a section devoted to recommendations for Sustainability. Options which are considered and discussed range from total client control and operation to StarNET management and ongoing service, the recommended approach being determined by local client priorities and resources.

Summary

The StarNET system is a communications network approach to managing the scoring, reporting, and storage of performance assessment data. The technology for each of the components exists and is improving rapidly. Implementing StarNET is a problem of integrating psychometrics and technology.

The central mission of StarNET is to provide high quality, rapid (perhaps overnight) scoring and reporting of assessments that require raters, carefully trained and monitored. No lesser goal will suffice if accurate information from seamless assessments are to be realized, and improvement in human performance is to occur.

Authentic, seamless performance assessment demands a multimedia capability, frequent ongoing assessments resulting in timely data for decision-making, and high quality, rapid scoring and reporting. These requirements serve as the parameters around which all StarNET applications are designed. There are four physical components to the StarNET system, each of which is customized for each client's unique situation:

- Local Multimedia Data Collection Center(s)
- Central Hub
- Rater/Scoring Interface Unit(s)
- Communications Network.

Since each StarNET application is intended to maximize the effectiveness of assessments in fostering the success of a local performance improvement program, the design for the StarNET system must be customized to each individual client's resources, priorities, and constraints. As such, local design factors require identification and incorporation into each StarNET application. In particular, four areas are explored in designing each StarNET system:

- Existing Assessment Program
- Scoring and Monitoring Considerations
- Uses of Assessment Data
- Infrastructure or Implementation and Sustainability.

StarNET specifications delivered to the client review options for each of the four components and customize the system to best meet a particular client's constraints and goals. Those specifications might be as simple as a flow chart integrating existing capabilities and suggesting staff responsibilities and time frames. On the other hand, StarNET specifications for a client could be a much more elaborate detailing of assessment development, hardware acquisitions, software modifications, and staff training.

Whatever the specific form StarNET takes for a given client, the system will deliver the following benefits:

- Fast turnaround of scoring results,
- Eliminating the need for move massive amounts of paper from place-to-place for scoring,
- Eliminating the need to pay travel expenses of raters,
- Inexpensive storage of performance assessments & portfolios, and
- Central auditing of assessment results.

Other benefits of StarNET also accrue. For example, digitizing multimedia data and storing them on optical media provide a virtually permanent record of performance. Indeed, scanning student essays and transferring to optical disk is cost competitive with microfiche for document storage and retrieval. Another side-benefit is that the training program for raters can be used with little modification for teacher in-service programs. A third benefit is the usefulness of the communications network for a variety of other purposes, such as e-mail, when it is not being used for assessment functions.

A final feature of practical significance is that the StarNET design is modular in the sense that components are technology enabled; they can be replaced as they are superseded by newer and/or cheaper technology. For example the communication system can be upgraded without affecting the

other components. Similarly, the rater training and monitoring model is generic: changes or additions to the type of assessment can be made independently of other systems.