

StarNET

A Network for Managing Digitized Multimedia Performance Assessment Data*

by
R. Robert Rentz
R&R Solutions, Inc.

Introduction

At the 1993 AERA/NCME Annual meeting, sessions on performance assessment abounded. Many definitions of performance assessment were offered but clarity remained elusive. A definition advanced only partially facetiously was that performance assessment "...was anything that was not multiple choice."

One reason for the practical popularity of multiple-choice tests for large-scale assessment is that they can be machine scored (i.e. we can scan those little answer sheet bubbles). Regardless of how performance assessment is defined, one characteristic that performance assessments seem to have is their use of raters in scoring rather than the use of machines for scoring. In performance assessments we must rely on human observers making judgements. Purchasers of large-scale assessments have coasted along for years on the unchallenged assumption that multiple-choice equals cheap while ratings equal expensive. Moreover, vendors have used this assumption to their own advantage, defensively avoiding innovation and stockpiling profits.

Our experience with writing assessment over the last two decades offers insights worth noting. In many large-scale writing assessments, examinees write one or more pages which are shipped to a central scoring site, sorted, manipulated and packaged, then rated by one or more raters, who have been trained by one or more scoring directors and managed by table leaders. Raters are retrained after it is discovered they have drifted from the standard and we pray the same scoring director is around the next year. Results are compiled, aggregated, lasered and shipped back three months later at a cost of \$2 per examinee (assuming two raters and 30% third rater resolution). I have sufficient evidence from my own experience that the direct cost of one rater rating one paper is no more than \$0.15. -- 7.5% of the total cost! The currently cumbersome and expensive methods used in many writing assessments should not be propagated as performance assessments increase in variety and scope. New options should be entertained that focus on rating quality and timeliness. From a practical perspective, if performance assessment is to achieve its potential, efficient and cost effective ways of managing performance assessment data must be developed.

* An earlier version of this paper was produced May 1993 and ©1993 by R. Robert Rentz. All rights reserved.

The lessons businesses have been learning during the last decade are also worth noting. Trading partners (vendors & customers) order goods and make payments using EDI (electronic data interchange) methods; just-in-time inventory and manufacturing components arrive within hours, not months, of their need. The argument is not that speed decreases costs (although it sometimes does), it is that speed enhances competitiveness (i.e. effectiveness). If vendor A allows you to order at the last possible moment and delivers the goods just before you need them, while vendor B requires three weeks advanced ordering and delivers when he's good-and-ready, will you buy from A or B? If customer A pays the moment the receipt of goods is verified, while customer B cannot process your invoice in less than 30 days, to which customer will you give the more favorable terms? This is not to say that cost is not important but it is no longer all important.

Those educational administrators who have been accepting assessment results months after students have been tested and who believe this information serves useful student and teacher purposes are not being well served. Such belated information might serve some accountability and summative evaluation purpose, but its value for helping a teacher teach a student is at best doubtful. Testing students in February, scoring papers in April and May, and distributing so-called diagnostic student results, in time for Fall teacher in-service is of limited value. Such information is simply not timely. We need high-quality, seamless, authentic assessments of students' and teachers' performance and we need the results now! Although cost must be considered, for feedback to be effective, quality and speed should be the driving forces.

Description of StarNET

Imagine 21st Century assessment using 21st Century technology.... 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 students' 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."

"Cool!" says Peggy.

Les takes Peggy's student's cassette tape of his Spanish speech, walks into the assessment center and sits in a little booth, surrounded by computer peripherals. The booth is called a Local Digitizing Kiosk. 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 scanner 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 student's writing results and a report evaluating his videotaped 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. It's great to have good, relevant information so fast!

Meanwhile, following Les's input of his data, the StarNET system has created several multimedia database record, tagged with the necessary codes, and the system has begun transmission of the database records to the Central Hub. The Central Hub operates much like an online transaction processing system, identifying the type of record coming in, storing it for archiving and passing that record to the Remote Rater Unit for scoring.

The Central Hub software identifies which Remote Rater Unit will receive each multimedia record for scoring. The identification process takes into account schedule, work load, rater calibration (using multifaceted Rasch models -- important to ensure scoring equivalence) and any other factor required by the particular assessment object being scored. In our example Les's video record is sent to two raters one located in Hawaii, the other on the Island of Majorca off the coast of Spain (a retired ETS employee). The Central Hub software selects six raters in California to receive subsets of the 25 essay records. Once selected, those records (along with any other records) are sent to the Remote Rater Units.

A Remote Rater Unit might be located anywhere, and the rater doing the ratings could be working-at-home, a member of a small group, or even another teacher at CVHS, moonlighting for extra money. Numerous possibilities exist for setting up Remote Rater Units and compensating individual raters. In fact, different raters might have different compensation plans.

When a rater completes an assigned task the results are transmitted back to the Central Hub, compiled and stored. Reports are prepared and e-mailed to Les Reed's and Peggy Murray's electronic mailboxes.

As the above example illustrates, StarNET is applicable to those classes of examinee responses that can be digitized and transmitted electronically. The examples of such responses illustrated here include: scanned images of a student's essay; video tape of a teacher's demonstration of teaching competencies; an audio record of a foreign language speech. Other examples might be photographs of a science project or a computer program written directly on another computer and transferred via floppy disk to the Local Digitizing Kiosk. 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. Thus the prevailing technology controls the applications.

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. While speed is important, other benefits of StarNET 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 (see below) 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 are generic: changes or additions to the type of assessment can be made independently of other systems.

StarNET is composed of four major components:

Local Digitizing Kiosk- Collects and digitizes performance objects

Central Hub-- Manages rater training, assignments, monitoring and reporting

Remote Rater Unit- Scores performance objects

Communications Network Moves the digitized objects between components

The figure below illustrates these components and their relationship; they are described in more detail in the sections that follow.

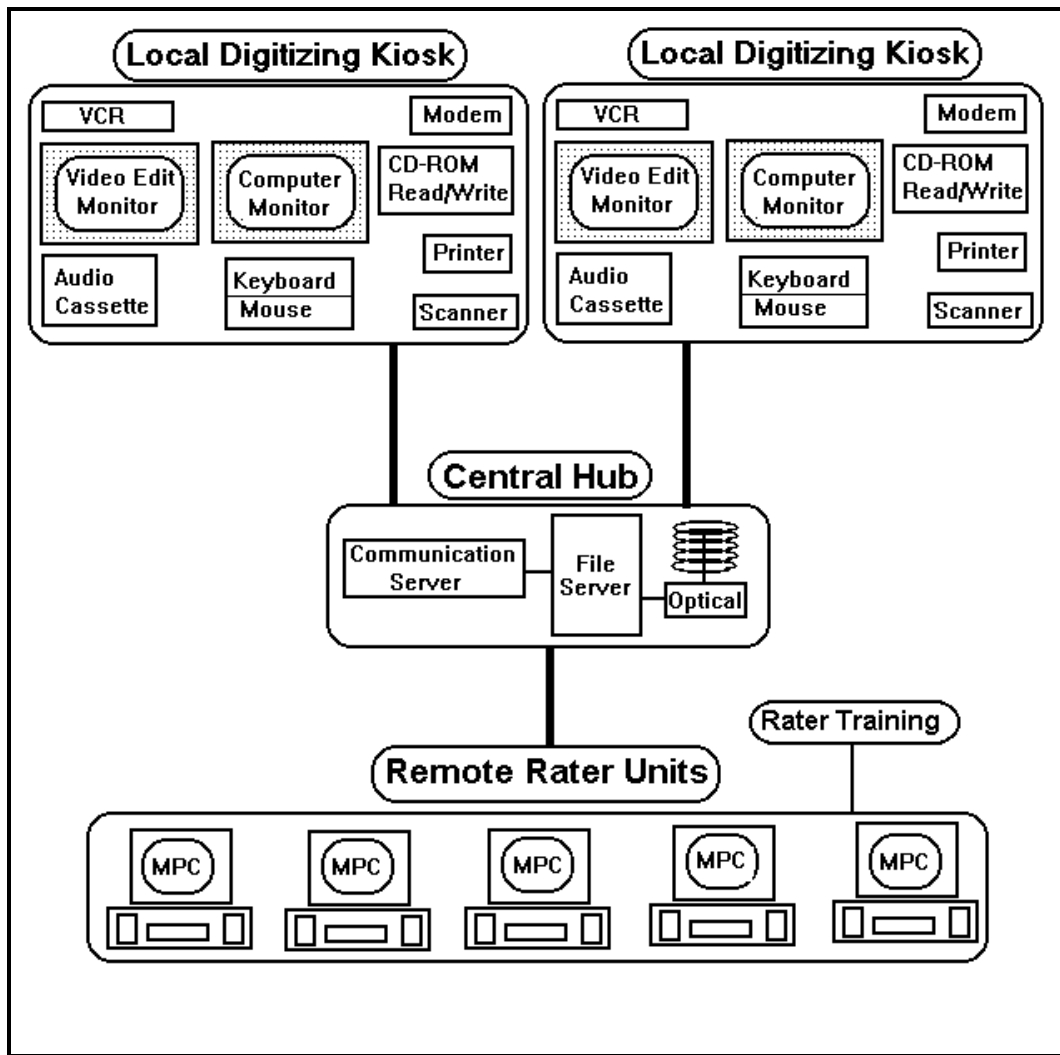


Figure 1: Components of the StarNET System

Local Digitizing Kiosk

Description & Function

A kiosk is a little booth; a Local Digitizing Kiosk might occupy floor space of about 5' by 5' and be about 5' high. I'm sitting at something similar, even as I write this paper; a "U" shaped work surface, with a monitor and keyboard in front of me. To my left is a document scanner; to my right a laser printer. I don't have a video or audio transfer unit connected to my system but I could. All the technology exists today to build a custom Local Digitizing Kiosk that will do the things I write about here.

A Local Digitizing Kiosk contains a computer and the peripherals necessary to transform assessment data into digital form. The type of data that can be digitized is dependent on current technology, and, thus, we can predict with confidence that peripherals will get better, faster, smaller and cheaper. The function of the Local Digitizing Kiosk is simple: it provides for inputting assessment data, transmitting that data to the next component of StarNET, and receiving the resulting reports.

One interesting feature of the Local Digitizing Kiosk is its potential mobility. The kiosk 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. Another interesting feature is the quantity of data the Local Digitizing Kiosk will have to handle. The types of data objects about which we are concerned require megabytes of storage.

Hardware Components

The hardware that might be included in a Local Digitizing Kiosk 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. The Canofile 250 is a document imaging system that scans 40 pages a minute.* Canon also makes the Zapshot, a digital color camera for taking still photographs. This camera's output goes to a floppy disk that can be transferred directly to StarNET.

The Local Digitizing Kiosk incorporates players for both audio and video tape that are connected to the PC through expansion boards that capture, digitize and compress these assessment objects. Components could be added or replaced as the technology advances.

Software Components

The software for the Local Digitizing Kiosk is straightforward in purpose, and, for the most part, software supplied by the 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.

* The inclusion of specific products is not intended as an endorsement, rather an example.

Central Hub

Function

The overall role of the Central Hub is management. Central Hub software receives the multimedia database records, stores them, determines which Remote Rater Unit will score them and routes these records to that Remote Rater Unit.

It then receives the resulting ratings and incorporates them into a report which is sent back to the sending Local Digitizing Kiosk. 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. This topic is discussed in a section below.

Hardware Components

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 would be small and the physical location arbitrary, dependent only on the Communications Network.

Software Components

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 must necessarily be 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 there is no reason a library of reporting options could not be provided.

Remote Rater Units

Function

The sole function of a Remote Rater Unit is to judge examinee performance, that is, to rate the assessment object. A Remote Rater Unit consists of a trained human observer, sitting before a multimedia PC, connected to the Central Hub through a high speed data link (probably a telephone line). The physical location of the Remote Rater Unit is arbitrary, it can be anywhere in the world. With standardized training and monitoring provided by the Central Hub, 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, timezones or a nine-to-five schedule. Quality ratings can be delivered like business inventories, "just-in-time."

Hardware Components

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 objects 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 object is large, that object only needs to be transmitted one-way. Only the ratings themselves have to be returned.

Software Components

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

Function

The function of the Communications Network is to move digitized assessment objects (i.e. multimedia database records) between the other 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 objects to well trained, human observers, to report their evaluations and to do this task fast.

Hardware and Software Components

I am less certain about the specific hardware and software components of the Communications Network than the other parts of StarNET. This area is undergoing rapid development and what is available today can be outdated six months from now. High speed modems are getting faster and cheaper. Wide area network services are proliferating. Applications like video conferencing and EDI systems are demanding larger capacity and faster communications services. The dilemma is this: by the time the other components of StarNET are designed and ready for implementation, the telecommunications field will be different. It is prudent to conceptualize the Communications Network component of StarNET a bit loosely for the moment.

Yet, it is likely that the first implementations of StarNET will use two kinds of distribution between the Central Hub and the Remote Rater Units. One would be batch, where a number of assessment records would be stored on CD-Rom (6000 pages of student essays per 5.25" platter) and delivered overnight to the Remote Rater Unit. The other method would transmit a single record at a time to the Remote Rater Unit over high-speed modem lines.

The communications link between the Local Digitizing Kiosk and the Central Hub will require a different strategy. In this case the volume of data is greater. The technology might include satellites, etc.

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.

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 we want raters to do is examine a particular assessment and to assign the correct rating. To do that they have to practice that specific task. They first 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 could use the same equipment that raters use for rating. Training content can be produced, published on CD-ROM, along with pre-rated carefully chosen examples. This model of training-by-example is outlined in Appendix A.

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

The content of a training program would be specific to a particular assessment and customized to local standards. The added benefits of creating a individualized, computer based, training program include the potential for use in a staff development program for teachers; and the necessary equipment would already exist at the Local Digitizing Kiosk.

Summary

In this paper I have presented the concept of StarNET, a wide-area network approach to managing the scoring, reporting and storage of performance assessment data. The technology for each of the components exists today and it is improving rapidly. Implementing StarNET is a problem of integrating psychometrics and technology. The goal is to produce high quality information, fast.

Appendix A

Training Program Content Structure (T-Shell)

Multimedia Description of

Assessment content
Scoring dimensions or categories
Scoring scale or rubric

Annotated Examples (Anchors or Models)

Model objects for each score point (dead-solid-perfect)
Explanation or rationale (text or voice)

Practice and Qualifying Objects 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 Objects

Objects used as "check-sets" to monitor raters

Since performance assessment requires human raters making judgments, rater training is the most important component to ensure quality measurement. The above figure outlines a generic model of a computer-based, training-by-example program. Presented as an interactive multimedia program on CD-ROM, such training would be individualized, effective and relatively inexpensive. The introduction could describe the assessment to be rated as a way of orienting the rater trainee and in a multimedia environment might be a short video clip. This introduction could be followed by an explanation of the scoring scale and rubric using graphics, video, audio clips together with annotated examples.

The annotated examples should be presented in exactly the same format in which assessment objects will be presented to raters for rating. These are the anchors, models or example objects chosen or constructed to clearly represent and define the score scale. As such these examples should clearly exemplify particular score points as defined by the scoring rubric -- *dead-solid-perfect*. As in the case of written essays, one paper might not fully reflect the richness of the rubric defining one score point. Nevertheless, a single example at each point on the score scale serves an important function

as an elaboration of the definition of that scale. The anchor/model/example renders concrete the inherently abstract definition of the score scale points.

Explaining why these example objects are quintessential representatives of their score point is the function of annotations, a function similarly performed by a talented human trainer. The annotations could be presented in pop-up windows, as audio clips or in other multimedia modes. The value of substituting these annotations for interactions between the talented trainer and the trainees lies in the standardization of the information received by all raters. Good teachers or trainers cannot help but customize their instruction to their individual 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.

The program description could 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. Regardless of the media on which it is presented, this description is the common orientation received by all raters.

After the score scale is defined with both abstract definition and concrete example, practice objects are presented to the trainee rater. The trainee examines and rates the object 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 program informs raters of the correct answers and provides justification for that answer. After a predetermined level of accuracy on the practice objects, a set of calibration or qualifying objects is presented; a specific standard of accuracy is required on these objects before a rater is allowed to rate real assessment objects.

Assuming a trainee *passes* the qualifying set of objects and is allowed to rate *real* objects, periodic monitoring of that rater's performance is desirable. In writing assessment, papers presented to raters for this purpose are often called *check-sets*. These objects 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 objects are important ingredients in monitoring rater performance. Whether these *re-calibration* objects are stored with the other training content or presented to raters as if they were real objects, is a matter of specific program design requirements. Nevertheless, it is significant to note that objects used as *check-set* objects, within the digitized *StarNET* system, have the physical appearance of any other *real* object. Such is not the case for check-set papers often used in typical writing assessments that are photocopies and clearly distinguishable from *real* papers.



SUMMARY OF *STARNET*[™] COMPONENTS

COMPONENT	MULTIMEDIA DATA COLLECTION CENTER	CENTRAL HUB	REMOTE RATER UNIT	COMMUNICATIONS NETWORK
FUNCTION	<ul style="list-style-type: none">• transforms assessment data into digital form	<ul style="list-style-type: none">• managing• routing• reporting	<ul style="list-style-type: none">• judges evaluate performance	<ul style="list-style-type: none">• moving digital objects between MDCC-CH-RRU
HARDWARE	<ul style="list-style-type: none">• document scanner• digital camera• audio transfer• video transfer	<ul style="list-style-type: none">• application server• communication server• optical jukebox	<ul style="list-style-type: none">• multimedia PC• Internet connection	<ul style="list-style-type: none">• telephone line• satellite• ISDN• intranet-LAN
SOFTWARE	<ul style="list-style-type: none">• data capturing• temporary storage	<ul style="list-style-type: none">• routing• reporting• storing	<ul style="list-style-type: none">• training• data entry	<ul style="list-style-type: none">• communications• network management• Internet
KEY FEATURES	<ul style="list-style-type: none">• stationary, or• mobile	<ul style="list-style-type: none">• small staff	<ul style="list-style-type: none">• variable schedule• work-at-home	<ul style="list-style-type: none">• interactive• e-mail• video conference