DOCUMENT RESUME

ED 361 182 SE 053 584

AUTHOR Lomask, Michal S.; And Others
TITLE The Safety Simulator: Scoring, Reliability and Validity of Interactive Videodisc-Based Assessment of Science Teachers.
INSTITUTION Connecticut State Dept. of Education, Hartford.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Apr 93
CONTRACT NSF-TPE-9154557
NOTE 31p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (Atlanta, GA, April 15-19, 1993).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Evaluation Methods; High Schools; *Interactive Video; Junior High Schools; Junior High School Students; Knowledge Level; *Laboratory Safety; Middle Schools; Science Activities; Science Education; Science Experiments; *Science Teachers; Secondary School Science; Secondary School Teachers; Simulation; *Teacher Behavior; *Teacher Evaluation; Teaching Methods; Test Reliability; Test Validity; Videodisks
IDENTIFIERS *Interactive Videodisks; Pedagogical Content Knowledge; *Subject Content Knowledge

ABSTRACT

An experimental Interactive Video Disc (IVD) assessment program, funded partially by the National Science Foundation, was developed to assess science teachers' knowledge of safe management of lab facilities and activities. The IVD program contains two phases: (1) panoramic view of the lab room, including safety equipment and storage of chemicals; and (2) simulation of a typical lab general science activity, performed by four middle school students. Examinees, consisting of beginning and experienced science teachers, were asked to identify and verbally respond to a variety of safety events which were simulated in the IVD program. Examiners' verbal responses along with the video contexts in which they occurred, were recorded by the IVD system and transferred to conventional video tapes which were later used for scoring. Reliability of scores for the four different categories of safety (physical facilities, Chemicals, Lab techniques, and students' behavior) examined by calculating the mean correlation coefficients among three scorers, was found to be moderate to high. Evidence for content and construct validity were studied through job relatedness analysis, safety expert judgment and known group performance comparisons. Videodisk images of the panoramic and laboratory stages and safety simulator scoring sheets are included. Contains 19 references. (Author/MDH)

# THE SAFETY SIMULATOR:SCORING, RELIABILITY AND VALIDITY OF INTERACTIVE VIDEODISC-BASED ASSESSMENT OF SCIENCE TEACHERS

By
Michal S. Lomask
Larry Jacobson
Laurin P. Hafner
Ginette Delandshere

Connecticut State Department of Education
Bureau of Research and Teacher Assessment

Paper presented at the Annual Meeting of the
National Association for Research in Science Teaching

Atlanta, Georgia, April 15-19, 1993



This research was supported in part by the National Science Foundation (NSF), Grant TPE- 9154557.
The opiniond expressed herein do not nrcrssar..y reflect the position or policy of the NSF

# The Safety Simulator: Scoring, Reliability and Validity of Interactive Videodisc-based Assessment of Science Teachers


Abstract

An experimental Interactive Video Disc (IVD) assessment program, funded partially by the NSF, was developed to assess science teachers' knowledge of safe management of lab facilities and activities. The IVD program contains two phases: 1) Panoramic view of the lab room, including safety equipment and storage of chemicals; 2) Simulation of a typical lab general science activity, performed by four middle school students. Examinees, consisting of beginning and experienced science teachers, were asked to identify and verbally response to a variety of safety events which were simulated in the IVD program. Examinees' verbal responses along with the video contexts in which they occured, were recorded by the IVD system and transferred to conventional video tapes which were later used for scoring. Reliability of scores for the four different categories of safety (*Physical facilities, Chemicals, Lab techniques* and *Students' behavior*) examined by calculating the mean correlation coefficients among three scorers, was found to be moderate to high. Evidence for content and construct validity were studied through job relatedness analysis, safet expert judgment and known group performance comparisons.



We would like to thank Mr. Earl Carlyon for his valuable assistance in the administration of the assessment. We would like also to thank the many students and science teachers who let us observe their classes. We have learned from each one of them.

# THE SAFETY SIMULATOR: SCORING, RELIABILITY AND VALIDITY OF INTERACTIVE VIDEODISC BASED ASSESSMENT OF SCIENCE TEACHERS

## Introduction

The Connecticut Teacher Assessment Center (CONNTAC) program for science teachers is currently being developed as part of the clinical assessment of professional knowledge of beginning teachers who are seeking their Provisional Teaching Certificate in Connecticut. Currently, beginning teachers are evaluated through classroom observations, aimed to assess a wide range of general pedagogical behaviors such as classroom management, questionings techniques and lesson planning. In contrast, the CONNTAC program is being designed to examine the integrated content-specific pedagogy, the unique "professional knowledge" of teachers (Shulman, 1986, 1987).

As a first step in establishing some of the critical content-specific pedagogical knowledge and job-related skills of science teachers, two sources of information were explored. One approach was a review of the relevant professional literature to determine *what teachers should know* (Gardner & Greeno, 1990; Leinhardt, 1990; Shulman, 1986, 1987; Strassenburg, 1989). This survey found that except for "walking on water" teachers needed to know almost everything: subject matter, general pedagogy, content-specific pedagogy, child psychology, communication, management, theories of cognition, learning disabilities and much more.

The second approach to establish content-bound pedagogy was to conduct a survey of practicing teachers to determine *what Connecticut teachers know and actually do* (Lomask and Ross, 1991a). In this survey, 20 beginning (less than 3 years of teaching experience) and experienced (more than 3 years of teaching experience) science teachers, who conducted lab experiments and other activities with their students, were observed, taped and interviewed. This study found that the use of hands-on lab activities in secondary science courses is quite limited. In most courses teachers were observed to offer hands-on lab activities only once a week. Many of these activities were confirmative in nature and only rarely an authentic science inquiry activity was observed. In addition, some of the lab activities that were observed were performed under unsafe conditions. Lack of basic safety equipment, as well as violations of OSHA (Occupational Safety and Health Administration) safety regulations were occasionally observed.

These observations and interviews with beginning and experienced teachers support what has been documented previously by other researchers (For example, Kaufman, 1992; Marsick and Thornton, 1988; Nagel, 1982; Reynolds, 1986). Teachers' lack of knowledge of safety issues limits the range of hands-on activities they offer to their students inside and outside the science classroom. Many teachers, especially those who are inexperienced, complained that lack of knowledge of safety management of lab activities is the main factor that limits their use of hands-on labs with their science students. Teachers' lack of prior safety training and fears of potential liability have increasingly led to a vast reduction in the number and quality of hands-on lab activities. Field trips (fear of lyme disease), blood typing (fear of AIDS), analytical chemistry (expensive waste disposal) and simple dissection in biology (concerns raised by animal rights organizations), activities that used to be the core of many science courses, have nearly disappeared from the science curriculum.

It is noteworthy that many of the beginning teachers in the study informed the researchers that they had not received any formal training in safety management in their teacher preparation courses. Lack of available safety coursework was confirmed by examining the syllabi of eleven Connecticut teacher preparation programs (Lomask and Ross, 1991b). This means that beginning science teachers enter the classroom with only limited training in safety issues. In many cases beginning teachers must learn these skills on their own, or if they are fortunate, receive support and supervision from a peer teacher, who guides them through the complex process of managing safe school science activities.

There can be little doubt that in order to encourage safe practice of hands-on lab activities in schools, the ability to manage a safe lab should be mastered by all science teachers, prior to their entrance to the science classroom. Consequently, to assure that all science teachers possess an adequate knowledge of lab safety, Connecticut chose to include the area of lab safety as part of the certification program.

## Development of the Pilot Assessment

Once lab safety management was identified as a initial assessment focus, the project team was confronted with how to best assess this knowledge. Traditional paper-and-pencil tests can only assess general knowledge of chemicals and safety procedures, but fall short in capturing the complexity teachers face in the lab. Such complexity often involves conducting a science activity with a class of up to 24

adolescents, anticipating and preventing accidents while continuously monitoring the appropriate performance of different lab procedures by students.

The primary goal of the project, therefore, was to develop a *practical, realistic, reliable* and *valid* assessment program that captures, through the simulation of an actual lab activity, the complexity of science classroom practice.

To simulate the complexity of a science activity in the laboratory this project used interactive videodisc (IVD) technology. As in any assessment, the specific knowledge and abilities to be assessed were first identified and then, an assessment simulation was designed to elicit critical teacher performances. Standards of performance and appropriate scoring guides were developed based on the performance of a representative sample of teachers in the IVD assessment, along with the review by a science committee composed of experienced CT classroom science teachers. Figure 1 summarizes the steps taken in the development of this assessment.

Insert Fig 1 about here

## The Safety Simulator Structure

The current version of the IVD safety simulator consists of two phases designed to represent what a science teacher needs to do to conduct a safe laboratory. The first phase simulates a walk around the laboratory to give the examinee the opportunity to check the various safety facilities in the room. The second stage simulates a general science activity performed by four middle school students.

The examinees, during the assessment, observe the simulation on a computer monitor. Using a mouse pointer, a microphone and keyboard, examinees are asked to identify and verbally respond to a series of safety violations in real time. That is, when examinees believe a violation is about to or has occurred, they are asked to stop the activity, focus on the hazard and offer an appropriate preventive or corrective action, as if they were the teachers in this specific lab. Examinees' responses (e.g., the errors that were identified, when the errors were identified, and the proposed action to be taken) are recorded by the IVD system for latter review, analysis and scoring. If the examinees feel uncertain about any event in the simulation, they may select a close-up view, which provides a closer and clearer

ability to get close-to-the-action in a typical science laboratory setting. Illustrations of the panametric and laboratory stages of the IVD assessment are provided in Figures 3 and 4.

## Safety Simulator Content

The IVD simulates a typical secondary school lab activity in which students conduct several simple experiments to learn about the characteristics of acids and bases (see Figure 2 for details of the lab activity). The simulation shows a gender-mixed group of four students, working in pairs, following the directions of the lab manual.

The lab activity: Acids and Bases

**Objectives:**

1. Determine whether a substance is an acid or a base.
2. Study what happens when acidic and basic solutions are mixed.
3. Study the reaction between metal and acid.

**Materials:**

| | |
|---|---|
| Unknown solution A | Test tubes |
| Unknown solution B | Glass stirring rod |
| 1% phenolphthalein solution | Bunsen burner |
| Small pieces of zinc | Ring stand with support ring |
| Red and blue litmus paper | Magnifying lens |
| 200-ml beaker | Wire gauze |
| Graduated cylinder | Funnels |

**Procedure:**

1. Obtain 5 ml of solution A in a test tube and 5 ml of solution B in another test tube. Label the test tubes A and B
2. Place a drop from each solution on a piece of blue litmus paper. Record the results.
3. Put two drops of phenolphthalein solution into each test tube. Record the results.
4. Combine the content of tube A and B into a 200-ml beaker. Record the results.
5. Place the beaker on a wire gauze on a ring stand. Light the Bunsen burner and heat the beaker until all the liquid has boiled away.
6. Let the beaker cool down and then examine the remaining residue with a magnifying lens. Describe what you see by words or drawings.
7. Obtain 5 ml of the two solutions in separate clean test tubes. Label as before. Add a few small pieces of zinc into each test tube. Record the results.
8. Summarize all that you have learned from the above experiments and give the report to your teacher.

Figure 2. Students' instructions for the lab activity

The students who performed the lab activity for the simulations were instructed and trained to purposely perform safety violations, thus creating the events to which the examinee could respond. Information about typical safety errors, hazardous chemicals and proper disposal procedures were based on OSHA's regulations (1992), Kaufman (1991) and Steere (1974).

Safety errors and concerns in the simulation were divided into four categories:

1. Physical facilities and structures
2. Storage, handling and disposal of chemicals
3. Use of improper lab techniques
4. Inappropriate student behavior in the lab

It is important to mention, that during the preparation and videotaping of the lab activity, students were instructed how to simulate both safe and unsafe procedures. For the sake of the students' safety, water and food coloring, instead of actual chemicals, were used and any "cuts" or "burns" were only staged for the simulation.

## Interactive Videodisc System (IVD)

To create a simulation which is sufficiently realistic to allow a candidate to assume the role of a middle school or high school teacher conducting a laboratory activity requires a system that can swiftly control video, audio and computer-based information. The key components of an interactive system include, a videodisc and videodisc player, a computer with video-graphics overlay card, a sound system and software.

### Videodisc

It was determined early in this assesment developmnent, that conventional videotape did not allow for realistic simulations since only very slow and imprecise searches of the tape are possible. Previous experiences of the project staff had shown that videodisc technology offered the speed and accuracy necessary for the proposed assessment.

In considering the use of videodiscs, there were two videodisc formats that were evaluated: the CLV (constant linear velocity) format, which allows for 60 minutes of full- motion video per side and the CAV (constant angular velocity) format, which

allows for 30 minutes of full-motion video per side. Although the CLV format allows twice the amount of video as the CAV format, it does not provide the ability to find and retrieve specific frames of video on the disc. By contrast, the CAV format, although limited in playing time, does allow for the random access of video. That is, used in conjunction with a computer, the CAV format has the capability to search and play any of over 54,000 frame of video. Consequently, this format was selected for the Safety Simulator since the CAV format afforded a combination of features necessary for this assessment project.

In addition to providing video information, the CAV videodisc's two audio tracks can be used to reproduce high quality stereo or to play in two different languages. For both the video and audio information, the reproduction quality is very high if the original quality of the recorded materials is very high.

The production of high quality video source materials generally requires the use of formats such as Hi-8, S-VHS, 3/4" or 1" videotape. For this project S-VHS format was used for the original videotaping. The original videotape was then copied to 1" tape, the format used during the editing process. The editing process involved copying selected video portions and close-up frames that would later be used in producing the videodisc. The editing process described above required the use of a high-cost commercial production facility. For future video production the project will utilize in-house S-VHS equipment, which will greatly reduce editing costs.

From the edited videotape, a CAV formatted glass "check videodisc" was then pressed. A "check disc" is traditionally produced prior to a final pressing of the master in order to preview the video prior to production of multiple discs. Typically, videodisc copies are then made from the master disc. For this project however, check discs were used in the safety simulator. While check discs are more fragile than master discs, check discs provided a high quality, low cost alternative to the high cost of producing a master disc, particularly when few copies are required.

Videodisc Player

For this project, a Pioneer LD-V8000 LaserDisc player was utilized. This player was selected for its .5 second access time, that is, portions of the video on any area of the disc can be searched and ready for playback within .5 second. Access time was important for the realism of the simulator, to allow the candidate to quickly move back and forth between selected portions of the video (wide views and close-ups). In addition to the speed of the laserdisc player, this model also included a serial port

connector which allowed for control by a computer, a necessity for the interactive simulation.

Computer System

The majority of IVD applications currently utilize either IBM compatible (MS DOS) personal computers or the Apple line of computers. When we first evaluated IVD systems, Apple computers required the use of two monitors. While one monitor displayed computer text and graphics, the other monitor displayed video information. At that time, overlay cards (hardware that would allow computer and video information to be displayed on a single monitor) did not exist. Since overlay boards did exist for IBM compatible computers, we chose to develop IVD systems using IBM compatible PCs. For this project, an IBM compatible 386 computer with the following components was used:

*Hard disk drive* - Due to the number and size of text and graphics files, and the large sound files that store candidates responses, a hard disk was necessary. For this project, a 40 megabyte hard drive was found to be satisfactory.

*Graphics card* - The graphics board used in this application was a standard VGA board.

*Overlay cards* - For IVD systems that display computer text and graphics on the same monitor with video information it is necessary to use an overlay card that can mix both forms of information. In this project we used a PSI VGAVISION I card produced by Processor Sciences, Inc. This card was compatible with the software used to control the IVD system and allowed for a variety of other video effects. One such effect used in this project was the compression and display of video in the upper quadrant of the monitor used in the response screen.

Sound System

To make the simulation more realistic, candidates responded to safety errors verbally. This approach was also advantageous in eliminating the need for candidates to type. As discussed earlier, using verbal responses also facilitated the scoring process. Since candidates responded verbally in the assessment, it was necessary to use sound digitizing software and hardware. For this project, Covox's Voice Master Key digitizing system was used. A preamplifier and external speakers were also used to playback audio from the laserdisc. When making scoring tapes, a

sound mixer was used to combine subjects responses with the audio from the videodisc.

Software

Researchers on this project developed programs for this simulation primarily using Microsoft's Basic 7.1. The programs control the playing of the video simulation, manage the examinees interaction with the simulation, record the examinees verbal responses and other data necessary for scoring, and produce scoring tapes.

For additional information about IVD specifications see Jacobson and Hafner (1991).

## Scoring

Development of Scoring

During the pilot assessment, examinees' verbal responses were recorded by the IVD system and transferred onto conventional VHS video tape. The scoring tapes included the following information:

Panametric Stage

- Video stills showing a specific safety concern.
- Audio (verbal responses) by the examinee describing his/her concerns regarding the specific event and suggesting corrective action.

Lab Activity Stage

- Ten seconds of video leading to a point in the video when the examinee responded.
- One still video frame (the point at which the simulation was actually stopped by the examinee).
- Examinee's audio description of the event and suggested preventive or corrective action to be taken.

This unique way of recording examinees' performances enabled raters to score the examinees' responses within the context in which they originally occurred. A committee of six experienced science teachers and safety experts observed the IVD simulation and, during rater training, identified all the observable safety errors in

the simulation. This was a necessary step since, although the errors embedded in the simulation were scripted in advance, students sometimes committed additional spontaneous errors during the taping of the simulation.

The committee examined the various responses and created scoring guides, with detailed benchmarks (examples of responses that represented different levels of performance) to guide the IVD scoring process. The levels refer to a rating of 0, 1, or 2. Zero was used when the examinees did not identify the error; 1 was used when the error was identified without a sufficient corrective action; and 2 was used when the examinee both identified and mentioned an appropriate corrective action.

Following identification of observable errors, the committee was asked to classify errors into four categories:

1) Physical facilities
2) Storage, handling and disposal of chemicals
3) Lab techniques
4) Students' behavior in the lab

Table 1 presents the percentage of agreement of safety committee members (N=5) by category of error.

Table 1
Percent of Safety Committee Agreement by Error Category

| Safety Errors Category | Number of Items | Percent Agreement |
|---|---|---|
| Physical facilities | 11 | 100% |
| Chemicals | 7 | 94% |
| Lab Techniques | 12 | 90% |
| Student Behavior | 9 | 86% |

The committee's next step was to establish criteria of performance, which was done through observations of examinees' audio and videotaped responses and through discussion of acceptable levels of performance. The committee used an expert judgment approach, using performance anchors, to assign a rating of 0, 1 or 2 to each

criterion, for each examinee. Sample pages of the scoring guide used by the raters are illustrated in Appendix 1.

With regard to the practicality of the scoring system, the length of time to score each tape ranged from 30 to 50 minutes with a mean of 38 minutes. This suggests that the videotape scoring system can be efficient, particularly since the administration of the assessment took approximately one to two hours per examinee.

## Evidence of Reliability

### Standardization of Administration

The use of IVD technology for the assessment provides for consistency of administration. All examinees, during the IVD assessment, received the same training and observed the same lab simulation with students making the same safety errors. The IVD system ensured that all examinees received uniform testing conditions.

### Reliability of Scores

Six science teachers, with knowledge and experience of safety in school laboratories, were trained as raters. The training included an overview of the program, identification of safety concerns and violations, classification of safety violations, use of scoring guides, discussion of exemplars of responses and scoring and calibration of scoring.

Calibration (i.e., anchoring raters to a common set of standards) included having the six raters observe the recorded responses of one prejuried examinee and to score her performance. Rater scores were compared and reviewed. After raters reached a satisfactory level of agreement, each individually scored the performance of 10-11 examinees. The design of the scoring study is shown in Table 2 and the analysis of the scores are presented in Table 3.

Table 2
Design of Scoring Study

| Candidates | rater 1 | rater 2 | rater 3 | ratter 4 | rater 5 | rater 6 |
|---|---|---|---|---|---|---|
| 1 | x | | x | x | | |
| 2 | x | | x | x | | |
| 3 | x | | x | x | | |
| 4 | x | | x | x | | |
| 5 | x | | x | x | | |
| 6 | x | x | | | x | |
| 7 | x | x | | | x | |
| 8 | x | x | | | x | |
| 9 | x | x | | | x | |
| 10 | x | x | | | x | |
| 11 | | x | x | | | x |
| 12 | | x | x | | | x |
| 13 | | x | x | | | x |
| 14 | | x | x | | | x |
| 15 | | x | x | | | x |
| 16 | | | | x | x | x |
| 17 | | | | x | x | x |
| 18 | | | | x | x | x |
| 19 | | | | x | x | x |
| 20 | | | | x | x | x |
| 21 | | | | x | x | x |

Note: The x's in Table 2 represent which candidates were scored by a particular rater. For example, candidates 11 through 15 were scored by raters 2, 3, and 6.

Table 3

Rater Means and Standard Deviations by Safety Categories

| Scale | Rater | N | Mean | Std Dev | Min-Max |
|---|---|---|---|---|---|
| Physical Facility | 1 | 10 | 0.54 | 0.43 | 0.08 - 1.23 |
| (Range = 0 - 2) | 2 | 11 | 0.29 | 0.29 | 0.00 - 0.85 |
| | 3 | 10 | 0.52 | 0.47 | 0.00 - 1.38 |
| | 4 | 11 | 0.33 | 0.39 | 0.00 - 1.38 |
| | 5 | 11 | 0.44 | 0.44 | 0.00 - 1.23 |
| | 6 | 10 | 0.58 | 0.43 | 0.00 - 1.08 |
| Chemicals | 1 | 10 | 0.58 | 0.31 | 0.22 - 1.22 |
| (Range = 0 - 2) | 2 | 11 | 0.51 | 0.46 | 0.00 - 1.44 |
| | 3 | 10 | 0.54 | 0.30 | 0.22 - 1.00 |
| | 4 | 11 | 0.28 | 0.28 | 0.00 - 0.78 |
| | 5 | 11 | 0.40 | 0.37 | 0.00 - 1.00 |
| | 6 | 10 | 0.51 | 0.30 | 0.22 - 1.11 |
| Lab Techniques | 1 | 10 | 0.99 | 0.39 | 0.33 - 1.57 |
| (Range = 0 - 2) | 2 | 11 | 1.06 | 0.28 | 0.52- 1.38 |
| | 3 | 10 | 0.77 | 0.31 | 0.33 - 1.19 |
| | 4 | 11 | 0.79 | 0.32 | 0.14 - 1.33 |
| | 5 | 11 | 0.97 | 0.30 | 0.48 - 1.33 |
| | 6 | 10 | 0.86 | 0.36 | 0.38 - 1.38 |
| Student Behavior | 1 | 10 | 0.90 | 0.23 | 0.50 - 1.38 |
| (Range = 0 - 2) | 2 | 11 | 1.01 | 0.27 | 0.50 - 1.38 |
| | 3 | 10 | 0.81 | 0.27 | 0.38 - 1.13 |
| | 4 | 11 | 0.87 | 0.21 | 0.50 - 1.38 |
| | 5 | 11 | 1.01 | 0.34 | 0.31 - 1.50 |
| | 6 | 10 | 0.86 | 0.28 | 0.44 - 1.38 |

Table 3 provides some information about rater tendencies such as leniency, strictness, and how much of the 0-2 point scale was used. Generally speaking, raters were remarkably consistent in their average ratings within categories. There were only two instances were raters appeared to differ with respect to their mean ratings. Rater 2 on physical facilities category and rater 4 on the chemicals category assigned lower average ratings than the other four raters.

Overall it was found that the physical facilities and chemical categories resulted in lower average scores than the lab techniques and student behavior categories. The mean ratings for the physical facility and chemical categories were 0.45 and 0.47 respectively. By contrast, the mean rating for the lab technique and student behavior categories were 0.91 and 0.91 respectively.

Frequency distributions of the ratings on each item showed that a rating of 0 (error was not identified) was frequently used. On some items, for example, all raters assigned a score of 0 for all examinees. Such items may have been too difficult. Sources of difficulty may have been the items themselves or the way in which they were presented in the simulation. For example, certain errors may have occurred too rapidly or may have occurred at the same time as other errors. Further analysis of each item will be necessary before conclusions can be made.

In Table 4, mean correlation coefficients for three raters' scores were calculated. Correlations were produced for each pair of those three raters (e.g., 2 with 4, 2 with 5, 4 with 5) for a common set of examinees within each category. Those three correlations were then averaged to produce the means reported in Table 4.

Table 4

Mean Correlations Between Rater Triads for Each Category

| | Rater Triad | Mean Correlation | N |
|---|---|---|---|
| Physical Facility | 2 - 4 - 5 | 0.82 | 6 |
| | 1 - 3 - 4 | 0.97 | 5 |
| | 1 - 2- 6 | 0.96 | 5 |
| | 3 - 5 - 6 | 0.96 | 5 |
| Chemicals | 2 - 4 - 5 | 0.95 | 6 |
| | 1 - 3 - 4 | 0.92 | 5 |
| | 1 - 2- 6 | 0.88 | 5 |
| | 3 - 5 - 6 | 0.98 | 5 |
| Lab Techniques | 2 - 4 - 5 | 0.93 | 6 |
| | 1 - 3 - 4 | 0.85 | 5 |
| | 1 - 2 - 6 | 0.94 | 5 |
| | 3 - 5 - 6 | 0.99 | 5 |
| Student Behavior | 2 - 4 - 5 | 0.66 | 6 |
| | 1 - 3 - 4 | 0.43 | 5 |
| | 1 - 2 - 6 | 0.91 | 5 |
| | 3 - 5 - 6 | 0.93 | 5 |

With the exception of two mean correlations within the student behavior category, the correlations between raters (the inter-rater reliability) were high. Although based on a very small sample size these correlations provide promising evidence that the scoring process can result in reliable scoring.

While results of Table 4 can provide a general indication of reliability, another measure of reliability is the level of agreement between raters. That is, did raters assign the same score to the same examinee on a given item? To calculate the inter-rater agreement, the number of times each pair of raters assigned the same score on a given item was summed and the mean agreement for the three rater pairs within each category was calculated. The percent agreement was produced by dividing the mean agreement by the number of items in each category. The results of this analysis are presented in Table 5.

Table 5
Inter-rater Agreement by Categories

| Scale | Items | Min | Max | Mean | SD | Agreement |
|---|---|---|---|---|---|---|
| Physical Facilities | 13 | 9.3 | 13.0 | 11.9 | 1.03 | 92% |
| Chemicals | 9 | 5.7 | 9.0 | 7.6 | 1.15 | 85% |
| Lab Techniques | 21 | 12.7 | 20.3 | 17.2 | 2.01 | 82% |
| Students Behavior | 16 | 10.7 | 16.0 | 13.7 | 1.49 | 85% |

Results presented in Table 5 show high levels of agreement between raters. This provides further evidence of the reliability of the scoring process.

## Evidence of Validity

Validation is a set of activities that progressively clarify the meaning of scores on a test (Pecheone & Carey, 1989). Also, validation studies do not validate a test per se, but rather address particular interpretations or uses of the test (AERA, APA, NCME, joint committee, 1974, 1985). With these assumptions in mind, we tried to establish the content, construct and criterion validity of the IVD assessment as a measure of science teachers' awareness and knowledge of safety regulation and behaviors in the school science laboratory.

### Validity Evidence Through Job Relatedness

Simulation-based assessment is one of the highly recommended methods of performance assessment (International Congress of the Assessment Center Method, 1975, 1989). For a simulation to be valid however, it must demonstrate job relatedness. In this project, a key approach to ensuring the instrument's relationship to the job was to involve experienced science teachers in nearly all phases of development and scoring. Teachers recommended different science activities as a basis for the simulation, as well as brainstormed the types of errors students are likely to make. This process was useful to ensure that the simulation measures characteristics that are relevant and crucial for the performance of the science teaching job and lends support to the content validity of the instrument.

Validity Evidence Through Expert Judgment

The measures for effective teaching are not sufficiently understood and articulated to serve as a basis for teacher evaluation (Berliner, 1988). But the measures of safety in the school laboratory are well understood and better articulated. Therefore the task of content validation by expert judgment was a straight forward one. Twenty five middle and high school science teachers performed the IVD assessment and then responded to the following feedback questions:

a. How well did the activity portray a regular middle/high school activity?
b. Are you aware of any safety concerns that were not addressed by the IVD simulation? If yes, please specify them.
c. How well did your role in this assessment resemble your role as science teacher in the science lab?
d. Would you expect science teachers to be knowledgeable about the safety issues that were simulated in this assessment?

The following is a summary of the examinee's feedback regarding the IVD simulation:

a. There was complete agreement that the IVD simulation portrays a realistic picture of science activity in the school lab.
b. Teachers stated that the simulation covers most of the safety issues in school science. Several biology teachers suggested that biology-specific issues, such as dissection and sterilization procedures, were not addressed and should be added to the simulation.
c. All teachers agreed that inspecting safety equipment and managing students' safety behavior, through the IVD, is very similar to what they do each day in their own classrooms. Some of the teachers added that mo ring four students on the screen is almost as difficult as monitoring 24 students in an actual lab activity. Other participants commented that an actual lab is more complex than was portrayed in the simulation. All agreed that their role in the assessment was very similar to that found in the school science lab.
d. All teachers agreed that IVD simulation assesses knowledge and skills that should be mastered by all science teachers.
e. Many of the teachers, while acclaiming the innovative use of technology in this project, indicated they would like to see a more professional, "TV quality" video

simulation. Most of the participants also recognized some mismatches between the video and the corresponding close-ups.

Results of the expert's judgmer. process, described above, provides support for the content validity of the IVD assessment.

Validity Evidence Through Known Groups Comparisons

A comparison of the performance of examinees from known groups taking the IVD assessment was conducted to study the meaning and interpretations of the assessment results. In this study eleven experienced science teachers, six beginning science teachers (some of these beginning teachers had a previous experience with safety issues, gained through science related jobs held before teaching), and four English teachers were administered the IVD assessment and their performances were compared. Each examinee was then independently scored by three raters. The mean scores across raters were used to compare the performance of the three groups. Table 6 presents the results of comparisons between the different groups of teachers on the four safety categories.

Table 6

Performance on the IVD Simulation, by Science Teaching Experience

| Safety Category | Non-science Teachers (N=4) | Beginning Science Teachers( N=6) | Experienced Science Teachers (N=11) | F | p |
|---|---|---|---|---|---|
| Physical Facility | 0.10 | 0.29 | 0.66 | 4.60 | .024 |
| Chemicals | 0.27 | 0.49 | 0.53 | 0.87 | .436 |
| Lab Techniques | 0.47 | 0.95 | 1.05 | 8.97 | .002 |
| Students' Behavior | 0.63 | 1.03 | 0.95 | 4.20 | .030 |

Significant differences were found between groups for the physical facility, lab techniques, and students behavior categories. For the chemical category there was no significant difference between the groups. Duncan's post-hoc test of differences between means indicated that, for the physical facility category, the experienced science teachers scored higher than the non-science teachers. For the lab techniques and students behavior categories, both groups of science teachers scored significantly higher than the non-science teachers. These results suggest that this simulation can

distinguish between science and non-science teachers and support the discriminative power and construct validity of the assessment.

## Discussion

Scientific experimentation in middle school and high school is considered by experts and teachers alike to be an important component of science teaching. From observations and interviews with beginning and experienced teachers, as well as a survey of the literature (Kaufman, 1992; Marsick and Thornton, 1988; Nagel, 1982; Reynolds, 1986), safety management of lab activities in schools emerged as a critical component of hands-on lab activities in science classes. In the State of Connecticut, however, pre teaching training in safety management was determined to be absent. Consequently, as a means of ensuring that teachers develop skills in safe lab management, Connecticut plans to integrate assessment of lab safety management as a requirement for beginning science teachers' certification.

Pilot results of the safety simulation suggest that it can be a practical, realistic, and valid assessment of safety management skills in the school lab. Initial results suggest that the performance of examinees can be reliably scored. Although the relatively small number of examinees in the pilot study does not allow for drawing final conclusions, the evidence obtained has been extremely positive and has encouraged the State of Connecticut to continue this line of research and development as a component of the certification process.

Often one of the main concerns in performance assessment has been the low reliability of scores produced by scoring systems which rely on human judgment as the source of examinees' ratings. However, the scoring system that was developed for the safety simulation was found to produce highly reliable scores. This is due, probably, to the highly structured and analytical approach adopted for the scoring procedures. Further, by adding video segments as a context to the examinee's recorded verbal responses and by adding video still reminders to the scoring guide (see samples in Appendix 1), the scoring process was simplified and the scoring time was greatly reduced.

Initial findings and analyses also support the validity of the new assessment. Science teachers performed significantly better than non-science teachers on three out of the four categories of the assessment. Oddly enough, in the chemicals category there were no significant differences found between the groups. Possibly

the items for the chemicals category were too difficult and/or there were too few items on which to accurately assess this knowledge. The next version of the safety simulation will likely provide a wider sample of items from the chemical category.

Feasibility of Administration and Scoring of the Simulation.

The time necessary for the administration of the assessment was 1-2 hours for each candidate while the scoring time averaged 30-50 minutes. The estimated costs associated with a state-wide assessment would be approximately $4,500 for each IVD station and any additional costs associated with the scoring process. Number of IVD stations necessary for a stste wide assessment depends on the size and geography of the state. In a small state like Connecticut, for example, we have found that four IVD stations, located in four regional educational centers, will be more than enough for continious assessment of beginning teachers. Based on Connecticut's experiences with different modes of assessment, it is expected that administration and scoring of the IVD assessment will compare favorably with other modes of performance assessment.

Feedback from the Field.

Science teachers, as well as safety experts who had hands-on experience with the safety simulation found the assessment to be highly realistic. Most of the participants were complimentary and enthusiastically supported the use of this simulation for the assessment of beginning science teachers. Further, many of the participating teachers requested that the simulation be made available for in-service professional development in their schools.

Based on the positive reception by science teachers and the promising results from the pilot study, the Connecticut State Department of Education is considering the use of the safety simulation as a key part of the requirements for beginning science teacher certification.

## References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) Joint Committee, (1974, 1985). Standards for educational and psychological testing, Washington, DC: American Psychological Association.

Berliner, D., 1988. Proceedings of the National Governor's Association Symposium on New Teacher Assessment. Boulder, Colorado.

Gardner, M. and Greeno, J. (Eds.), 1990. Toward a Scientific Practice of Science Education, Lawrence Erlbaum Associates.

Jacobson, L. and Hafner, L. P., 1991. Using Interactive Videodisc Technology to Enhance Rater Training. IPMA Assessment Council, Annual meeting, Chicago.

Kaufman, J., 1991. Laboratory Safety Workshop. Curry College, Milton, MA.

Kaufman, J., 1992. Leadership in safety: It's time to wake-up America. Speaking of Safety, Spring 1992.

Leinhardt, G., 1990. Capturing craft knowledge in teaching. Educational Researcher, 19, 18-25.

Leonard, H., W. (1992). A comparison of student performance following instruction by interactive videodisc versus conventional laboratory. Journal of Research in Science Teaching, 29, 93-102.

Lomask, M., Jacobson L. and Hafner L. (1992). Interactive Videodisc as a tool for assessing science teachers' knowledge of safety regulations in school laboratory. Paper presented at the annual meeting of the National Association of Research in Science Teaching, Boston.

Lomask, M. and Ross, C., 1991a. Teachers' ways of knowing - information from observations, interviews and questionnaires to science teachers (Connecticut State Department of Education, Internal report).

Lomask, M. and Ross, C., 1991b. Survey of Content of Science Teacher Preparation Programs in Connecticut (Connecticut State Department of Education, Internal report).

Marsick, J. D. and Thornton, S. F., 1988. Science teacher safety survey. Journal of Chemical Education, 65, 448-449.

Nagel, M., 1982. Lab magic and liability. The Science Teacher, 59, 31-33.

Pecheone, R. and Carey N. 1989. The validity of performance assessment for teacher licensure: Connecticut's ongoing research. Journal of Personnel Evaluation in Education, 3, 115-141.

Reynolds, R. F., 1986. Safety is your department. The Science Teacher, 63 (10), A242-A247.

Shulman, L. S., 1986. Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4-14.

Shulman, L. S., 1987. Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57, 1-22.

Strassenburg, A., A. (1989). What teachers need to know to teach science effectively. In Competing Visions of Teacher Knowledge: Proceeding from an NCRTE seminar for Education Policymakers, Vol. 1: Academic Subjects.

Steere, N. S.(Ed.), 1974. Safety in the Chemical Laboratory. Reprinted from Journal of Chemical Education.

Figure 1

Development of the IVD Science Safety Assessment

**Purposes of the Assessment**

- Certification of Science Teachers
- Assessment of Content Pedagogy
- Application of IVD Technology for Assessment

**Knowledge and Abilities to be Assessed**

- Knowledge of Safety Regulations
- Ability to Observe and Identify Safety Errors and Hazards
- Ability to Take Preventative and Corrective Actions with Errors, Hazards, and Accidents

**Development of Simulations to Elicit Knowledge and Abilities**

- Identification of Appropriate Laboratory Activities
- Identification of Appropriate Laboratory Errors and Hazards
- Design of IVD Process to Allow Examinees to Respond in "Real-Time" During the Laboratory Simulation

**Development of the Scoring System**

- Identification of Errors and Preventative and Corrective Actions by Science Teachers
- Examination of the Role of "Time" in the Scoring Process
- Preliminary Development of Branching Simulations

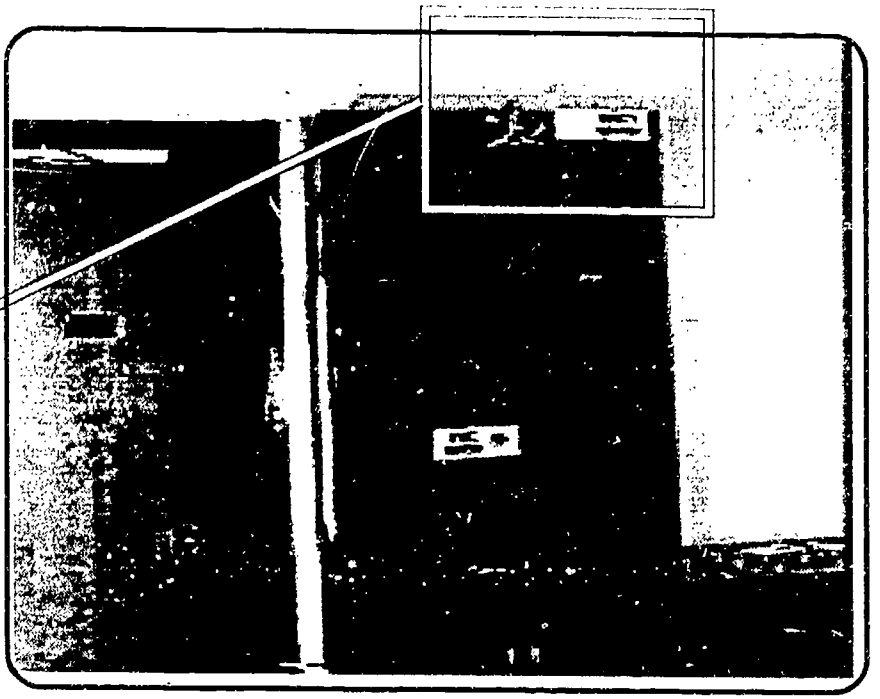
Figure 3

Panametric Stage

**FULL-MOTION WIDE-VIEW**

Examinee observes slow moving wide- view of the preparation area prior to the start of the lab. The examinee may stop the scene at any time for a closer look.

Using a mouse, the examinee points to an area of the scene in which they would like a close-up.



**STILL-FRAME CLOSE-UP**

A still-frame close-up view of the scene is then provided.


**VERBAL RESPONSE**

The examinee enters a voice response which identifies the suspected error is and suggests an action or solution.

**FULL-MOTION WIDE-VIEW**

The scene then reverts back to the slow moving wide-view of the preparation area from the point it was stopped.

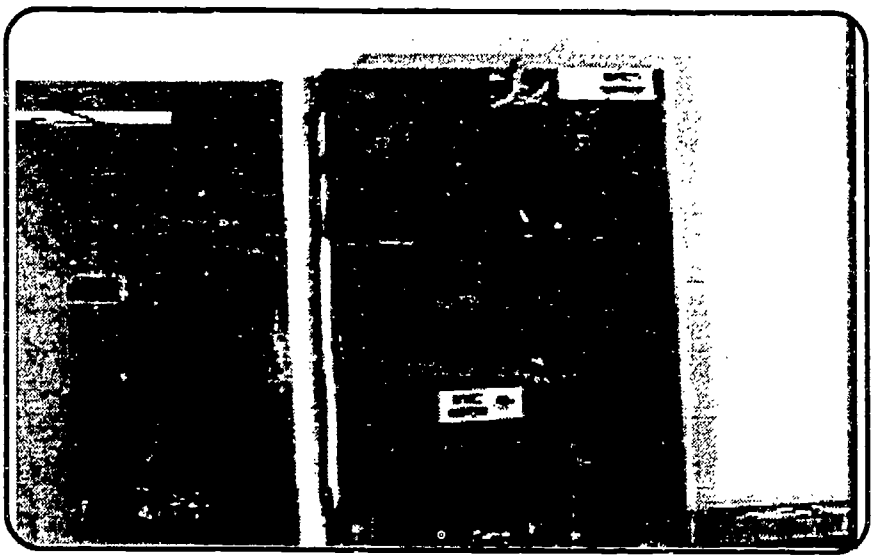
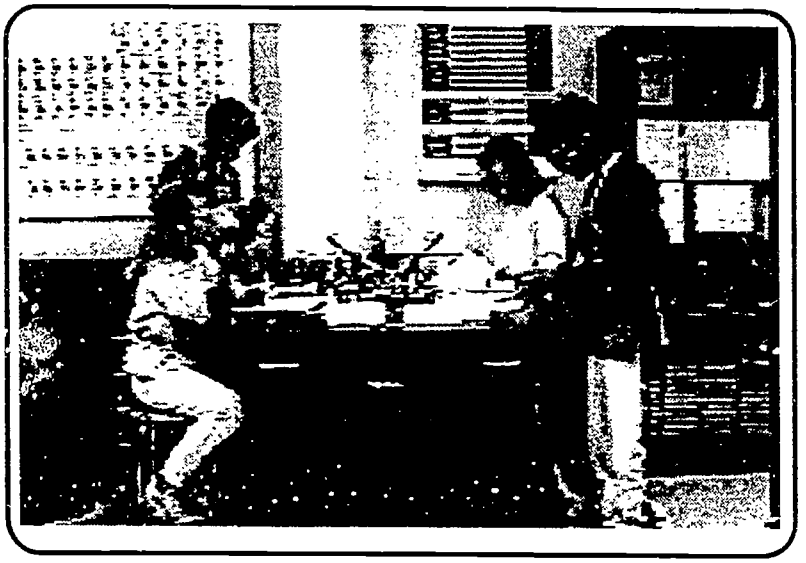


Figure 4
Laboratory Stage

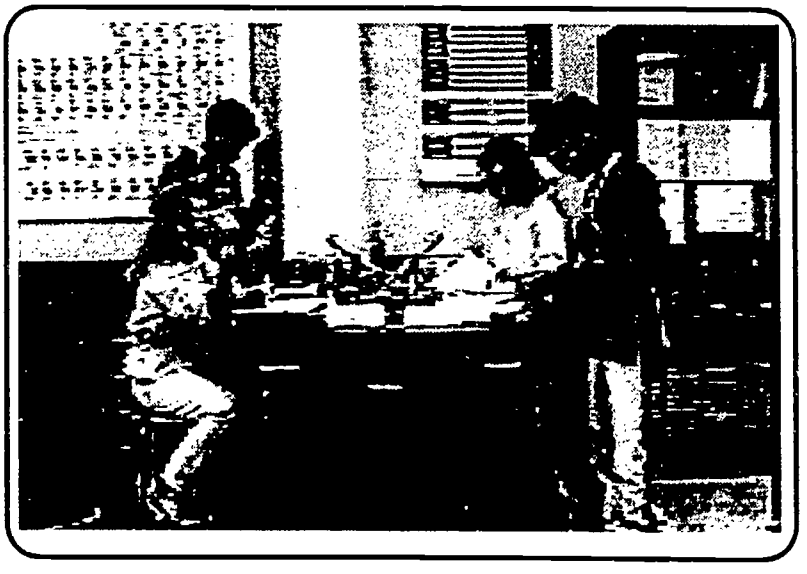
**FULL-MOTION WIDE VIEW**
Examinee observes lab and selects right or left work-station when a "closer look" is needed.



**STILL-FRAME CLOSE-UP**
If examinee selects the right side, a still-frame close-up of the picture is presented.

**VERBAL RESPONSE**
The examinee enters a voice response which identifies the event and suggests an action or solution.

**FULL-MOTION WIDE VIEW**
The scene then reverts back to full-motion wide-view and the laboratory continues from the point it stopped.

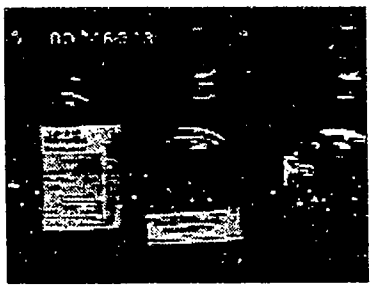


# The Safety Simulator - scoring sheet

## I. Panoramic view

Frames | Safety errors | Score

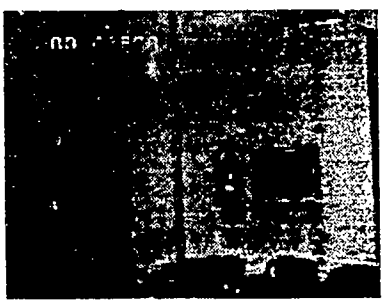
46943 **HOOD**

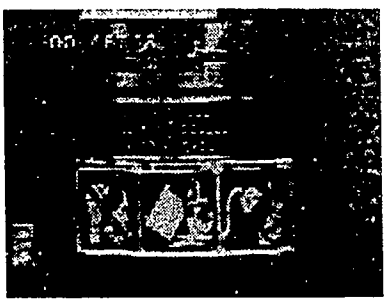


| | | | | |
|---|---|---|---|---|
| p | Material stored in hood | 0 | 1 | 2 |
| c | Carbon Disulfide in the hood | 0 | 1 | 2 |
| c | No warning label on acid bottle | 0 | 1 | 2 |

---

41500 **EXTINGUISHER** 46948

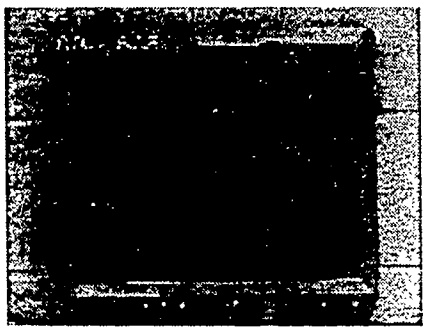




| | | | | |
|---|---|---|---|---|
| p | Wrong type of fire extinguisher | 0 | 1 | 2 |
| p | No sign for the extinguisher | 0 | 1 | 2 |
| p | Chairs are blocking access to the extinguisher | 0 | 1 | 2 |

---

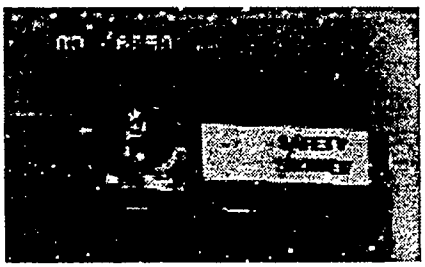
46954 BLANKET



p No written sign [0] [1] [2]

---

46959 SHOWER and EYE WASHER



p Chain is up [0] [1] [2]

---

46959 GOGGLES

p Wrong storage of goggles [0] [1] [2]

---

43400 CHEMICALS

c Chemicals are stored in an open shelf classroom [0] [1] [2]

c Ether is dangerously out of date [0] [1] [2]

c Chemicals compatibility is not kept [0] [1] [2]

## II. Lab Activity

### 1. Preactivity

00010- 00040

l Students working sitting on stools [0] [1] [2]

37749

b Books on bench, bags on floor (either one) [0] [1] [2]

---

38091

47039

b Girl's hair is not tied back [0] [1] [2]

b Boy's cloths are inappropriate for lab [0] [1] [2]

---

### Lab Activity

02251

p. Goggles storage, again [0] [1] [2]

38099

37775

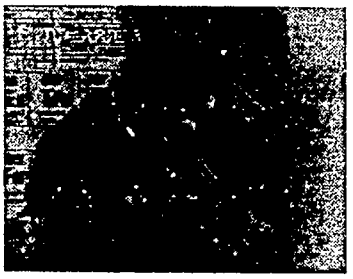
49004

1 Quinn pulls stopper and smells substance 0 1 2

1 Dumping unknown chemical down the sink 0 1 2

1 Andy is fixing his goggles 0 1 2

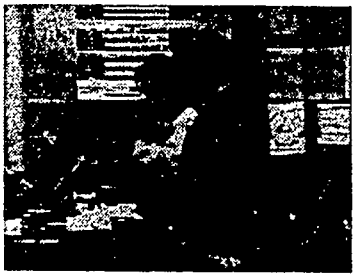
380043



1 Andy, mouth pipetting 0 1 2

---

37785



1 A paper towel is used to clean up spilled chemical 0 1 2

---