Direct, Indirect, and Controlled Observation and Rating Accuracy

Ann Marie Ryan, Diane Daum, Todd Bauman, Michele Grisze, Karen Mattimore, Todd Nalodka, and Sheila McCormick
Bowling Green State University

Videotaping of assessment center exercises has become an increasingly common practice, yet little is known about the impact of video technology on rating accuracy. This study compared ratings of a group discussion made after live observation (direct), after viewing a video (indirect), or after viewing a video with opportunities to pause and rewind (controlled). Results indicated some differences in observational accuracy but not in rating accuracy. Implications for the use of video technology in assessment centers are discussed.

From 1986 to 1991, the National Football League (NFL) used instant-replay officiating for clarifying situations where officials had difficulty in observing or observers had differing opinions on what had occurred. The opportunity to look carefully and repeatedly at a segment of behavior on the playing field was seen as an aid for situations where the observation ability of human judges was limited. "Replay officiating" would be an ideal tool for examining work performance, given the limited observational capacities of raters in the workplace. Unfortunately, in many work settings, the use of video to capture job performance is impractical or impossible. One rating setting where videotaping of performance is possible and has become widespread is the assessment center.

Bray and Byham (1991) noted that videotaping is being used to replace direct observation of assesses by assessors, allowing the evaluation of candidates at remote locations and the use of "professional assessors." Other stated advantages of videotaping assessment center exercises include a reduction in assessor fatigue that occurs from continual attention to the observation of others (Lepard, Edgemon, & Burns, 1990), enhanced validity by increasing the number of assessors who observe assesses (Pyne, Bernardin, Benton, & McEvoy, 1988), a reduction in the number of assessors required, enhanced credibility of assessment, the ensured comparability of judgments, lower costs (Olshfski & Cunningham, 1985), the ability to review videotapes when there are assessor disagreements in ratings (Cunningham & Olshfski, 1985), and the availability of the videotape for legal review (Frank, Bracken, & Struth, 1988).

This study was prompted by the implicit assumption of improved rating processes by using videotapes of assessment center exercises rather than viewing assesses performance live. NFL statistics showed reversals of official decisions ranging from 9.9% to 15.8% (M = 12.53%) of plays closely reviewed by the replay official in the 1986–1991 seasons (Shuster, 1992). The NFL decided that small gains in observational accuracy were offset by delays in action and missed fan support. The assumption that videotaped assessment centers (or other work situations) might be observed more accurately than concurrent observation and that those gains in accuracy would be substantial enough to warrant regular use of videotaping were the foci of this study.

Information-Processing Theory

The usefulness of videotaping can be discussed in terms of theories of information processing and social judgment. Cognitive models of the evaluation of others (e.g., DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980) have presented the performance-evaluation process as consisting of the observation of behavior and the encoding, storage, retrieval, and integration of information. The use of videotapes may have effects at each stage of the rating process.

First, the observation of behavior is affected by both...
conscious and unconscious information-acquisition strategies (Murphy & Cleveland, 1991). As a conscious strategy, an assessor viewing a videotape may opt to be more narrow in his or her observational goals (e.g., seek information on a specific behavioral dimension), knowing that the opportunity to rewind will provide a second chance at observing in a different observational goal. Videotaping may lead to less attention overall, as the need to be vigilant in observation is less when one knows there is the capacity to replay and catch what is missed. At the unconscious level, videotaping of performance may affect the salience and the distinctiveness of behaviors (e.g., gestures may be more or less noticeable in a two-dimensional presentation or recordings may make tonal variations more or less distinct; e.g., Murphy & Cleveland, 1991; Taylor & Fiske, 1978). Olshefski and Cunningham (1985) reported that assessors had difficulty in judging the eye contact between assessment center participants from a videotape, suggesting that there might be subtle differences in the information provided by video versus direct viewing. Distractions from the nonrated (i.e., nontarget) individuals may be greater in the live situation than in the videotaped situation, thereby affecting observation.

Videotaping with the opportunity to rewind and pause may affect the encoding of performance information. In direct observation, the assessor must attend to a steady stream of behavior, which is likely to include redundant, ambiguous, and inconsistent information (Dipboye, 1987). The capacity to review segments of behavior may enable assessors to better encode what has occurred. Indeed, Hauenstein (1992) found that allowing participants to stop a videotape might reduce encoding bias because a rater could process a performance incident without being distracted by ongoing rater performance.

Videotaping has clear implications for the storage and retrieval of information, because the videotape provides a complete memory aid. Although assessors typically record observations as an individual performs them, the level of detail is not as great as what one may obtain with a video record to review. However, Murphy and Balzer (1986) cautioned that more detail in recorded observations does not necessarily relate to greater accuracy in evaluation.

Social judgment research also suggests that videotapes may affect a rater’s behavior. For example, viewing a videotape versus observing performance live may create differences in affect toward the ratee, which has been shown to influence evaluations (e.g., Cardy & Dobbins, 1986). Perhaps stronger affective reactions occur when assesses are physically present versus when assesses are presented on a screen. Differences in assessor feelings of accountability (R. G. Gordon, Rozelle, & Baxter, 1988; Longenecker, Sims, & Gioia, 1987) may also occur because the assesse does not see the assessor in the videotaped situation but does in the direct observation situation. In sum, models of information processing and social judgment research suggest that videotaping in and of itself may or may not affect the rater’s behavior, but the control provided by replaying should enhance the ability to process information.

Research Comparing Direct and Video Observations

Little research is available that compares direct observation with video observation. Lepard et al. (1990) conducted a study that compared live (direct) observation of a fact-finding exercise by six assessors with video observation by another six assessors, with one rewind allowed. Those viewing videotapes reported less fatigue in recording behaviors, less stress, increased confidence in accuracy of recordings, less fatigue in preparing reports, more confidence in report accuracy, and less difficulty in preparing reports, but no differences in perceptions of difficulty of recording observations. Assessors viewing a live exercise indicated assesses asked more questions as part of the exercise, but there were no differences in assessor recordings of the number of facts uncovered or the number of recommendations made by assesses. Lepard et al. also concluded that dimension ratings were comparable across the two conditions. Moore and Lee (1974) compared ratings of six actual interviewers in live interviews with those of six individuals viewing a videotape of the same interview (no indication was given as to whether these individuals could pause or rewind the videotape). They found that the videotape viewers rated candidates significantly higher than did live interviewers on 5 of 12 dimensions and on the overall rating. The small sample sizes in these studies preclude any definitive conclusions.

Support for the comparability of live and videotape ratings comes from Buckner (1984), who found that ratings of 22 assesses in three exercises were similar for live and video viewers. Buckner found low to moderate agreement in the rankings and in the correlations of dimension ratings; an average correlation of .43 (between the two groups) in overall summary scores was obtained. Buckner’s results provide weak support for the comparability of ratings across the situations: Within-group assessor variance might have overshadowed between-groups differences, and the accuracy of ratings was not considered. Thus, the few studies that have examined live versus video ratings are inconclusive.

Hypotheses

This study examined differences in rater accuracy among those (a) directly observing an assessment center
exercise, (b) indirectly (through videotape) observing without the ability to control the rate of observation (i.e., no pausing or rewinding), and (c) observing with control (i.e., videotapes with rewinding and pausing). Rating accuracy can be assessed with several different measures. Murphy (1991) noted that the choice of behavioral-accuracy (e.g., recognition of specific behaviors) versus classification-accuracy (e.g., ratings compared with true score) measures depends, in part, on the purpose of rating. In assessment centers for selection and development, both types of accuracy are important, as ratings are used for both decision making and feedback. Thus, both observational-accuracy and rating-accuracy measures were incorporated into this study.

Given that models of information processing suggest that there are benefits from rewinding and pausing and that videotaping itself affects observation and evaluation quality (albeit with a mixed view on directionality), we examined the following hypotheses:

**Hypothesis 1A**: Differences in observational accuracy will exist between those viewing a live group discussion (direct observation) and those viewing a videotape of the discussion (indirect observation).

**Hypothesis 1B**: Observational accuracy will be greater for those in controlled observation settings than for those directly or indirectly viewing a discussion.

**Hypothesis 2A**: Differences in rating accuracy will exist between those directly viewing a group discussion and those viewing a videotape of the same discussion.

**Hypothesis 2B**: Rating accuracy will be greater for those in controlled observation settings than for those directly or indirectly viewing a discussion.

**Method**

**Participants**

Introductory psychology students \( N = 179 \) participated for course credit. With this sample size, the power to detect a medium effect at an alpha of .05 for one-way analyses of variance with three conditions was .86. Seventy-nine percent of the participants were female and 21% were male, with 65% being first-year undergraduates and 35% being second-fourth-year undergraduates. The mean age of the participants was 19.2 years \( (SD = 2.42 \) years).

**Materials**

A group-discussion exercise was chosen that would enable the assessment of common assessment center dimensions and that would be relevant to the student participants. The group discussion focused on raising money to meet a debt incurred at a small college (Jaffe, 1971).

Discussion-specific behaviors were generated by two graduate students to represent high, medium, and low levels of six dimensions (drive and initiative, organization and planning, persuasiveness, communication skills, listening and sensitivity, and judgment and decision making). The behaviors were then retranslated by a group of five graduate student raters. Any behaviors not correctly retranslated were removed from the study.

Scripts were then developed for two target individuals, a man and a woman, who would be the focus of the assessment. The first target individual was given a script containing behaviors that reflected high levels of listening and sensitivity and of judgment and decision making, medium levels of organization and planning and of drive and initiative, and low levels of persuasiveness and of communication skills. The second target individual was given a script containing behaviors that reflected high levels of drive and initiative and of persuasiveness, medium levels of judgment and decision making and of communication skills, and low levels of organization and planning and of listening and sensitivity. The three other group members were not given specific dimension-based scripts to follow; however, they were directed to provide certain responses to the statements of the target individuals.

**Measures**

**Behavioral observations**. Observations recorded by participants during the discussion were evaluated by two raters who were unaware of the participants' condition. These raters tallied the number of good behavioral observations and discussed any disagreements to reach consensus. According to Gaugler and Thornton (1989), a good observation is one that is an actual behavior rather than an inference or an impression. We had hoped to compare the written observations with those behaviors actually present as a measure of observational accuracy. However, the number of good written observations was very low for one target; we opted to use other measures of observational accuracy, which are described below in the **Behavioral checklist** section.

**Behavioral ratings**. Participants completed a dimensional rating form for each target to evaluate six dimensions. The dimensions were rated on a 6-point scale, with 0 indicating no chance to exhibit the dimension and 5 indicating a high level of the dimension. True scores were developed by two expert raters (one professor and one graduate student) who viewed all three versions of the videotaped discussion (described below in the **Procedure** section) and the script multiple times (Borman, 1977). There were no cases of disagreement greater than 1 rating point; differences were resolved, and expert ratings were used as true scores for computing measures of rater accuracy. Four rating-accuracy measures were used: (a) elevation, the accuracy of the average rating of a rater across both targets and all dimensions; (b) differential elevation, the accuracy of the average rating of a rater to each target across all dimensions; (c) stereotype accuracy, the accuracy of the average rating on a dimension by a rater across both targets; and (d) differential accuracy, an index of rater ability to discriminate among raters within dimensions (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

**Behavioral checklist**. Participants also completed 35-item observational checklists (one for each target) that included behaviors demonstrated by the target individuals as well as nonrelevant behaviors. This procedure was similar to that used by M. E. Gordon (1970). To ensure that these behaviors were present or absent as intended in all cases, two individuals (a gradu-
ate student and a professor) reviewed the videotapes. Eleven behaviors were present for the first target individual, and 15 behaviors were present for the second target individual; thus, 9 checklist behaviors were not exhibited by either target.

Observational-accuracy measures computed included hit rate (correctly identified as present); false-alarm rate (incorrectly identified as present); Pr (hit rate − false-alarm rate; range from −1 to 1), which indicated overall ability to distinguish between present and absent behaviors; and Br, which was a tendency to rate behaviors as present (false-alarm rate/1 − [hit rate − false-alarm rate]; range from 0 to 1, with numbers greater than 0.5 indicating a liberal bias and those less than 0.5 indicating a conservative bias; Borman & Hallam, 1991).

Controlled observation measures. In the controlled observation group, experimenters, observing through a one-way mirror, recorded (a) the number of times the videotape was paused by the participant and (b) the number of times the videotape was rewound. These behaviors were easily observable as the video monitor showed "pause" or "rew," respectively. Participants in the controlled observation condition were also asked to provide two ratings: using a 5-point scale ranging from 1 (not at all helpful) to 5 (very helpful), on how much rewinding and pausing helped them make more accurate ratings and to answer two open-ended questions regarding why they paused or rewound.

Procedure

The students participated in a 1 1/2 hr assessor training session in groups of 20 or fewer. This session followed the recommended content of Moses and Byham (1977) and included an introduction to the assessment process; dimensional definitions; classification and rating exercises; and practice in observing, recording, and evaluating behavior.

Immediately following the training session, participants either viewed a live group discussion (n = 51) or a videotape of that same discussion (n = 82) or were instructed to come back for individual sessions within the same week for an opportunity to watch the videotape (controlled observation condition; n = 46). To ensure adequate ability to observe, live and video discussions were not viewed by more than 20 individuals at once. Pilot testing had indicated that larger groups reported difficulty in hearing or seeing all group members. This required repeating the live discussion by the actors and videotaping on three separate occasions. Participants in the indirect and controlled observations were randomly assigned to view one of the three videotaped versions.

In all conditions, participants were instructed to watch the two targets and record observations while watching the discussion and then to evaluate behavior on the dimension rating forms after the discussion. After the rating forms were collected, participants were given the observational checklists to complete (one for each target individual). The checklists were administered after the dimension rating forms to eliminate the effects of potential prompting of nonrecorded observations, leading to changes in ratings.

Participants in the controlled observation conditions were observed through a one-way mirror. They were aware of the observation but not of the specific purpose of the observation. Participants were debriefed immediately following the session.

Results

Table 1 presents descriptive statistics for all of the dependent measures. Differential accuracy was related to stereotype accuracy; otherwise, the evaluation-accuracy measures were uncorrelated. Roach and Gupta (1992) demonstrated that accuracy components are relatively independent except for differential accuracy and stereotype accuracy. Their meta-analysis indicated a sample-weighted mean correlation of .33 between these components. Observational-accuracy measures were intercorrelated. Differential elevation was related to several of these measures.

There were no differences across the three viewing conditions in hit rate, false-alarm rate, and Br. There were differences in Pr, F(2, 176) = 3.92, p < .05, such that those in the controlled observation condition were significantly better at distinguishing behaviors as present or absent than were those in the indirect viewing condition (see Table 2). There were differences in the number of good observations recorded by condition, F(2, 176) = 3.32, p < .05, such that more good observations were recorded by those in the controlled viewing condition than by those viewing the discussion live. Thus, Hypothesis 1A was not supported, whereas Hypothesis 1B received limited support.

We conducted analyses of variance to examine if there were differences in rating accuracy by viewing condition (see Table 2). No differences in any of the measures were found. Thus, Hypotheses 2A and 2B were not supported.

Those in the controlled viewing condition paused more than they rewound, and they viewed pausing as more helpful than rewinding. There was considerable variability in the extent to which participants rewound (0–13 times) and paused (0–45 times) the videotape. As Table 1 shows, observational accuracy and the number of good observations recorded were unrelated to the amount of pausing and rewinding by those in the controlled observation group or to views on the helpfulness of these actions. Greater rewinding was related to lower elevation accuracy.

Responses to open-ended questions regarding why participants paused or rewound were fairly uniform. Participants said they paused to give them more time to record observations without missing subsequent action and that they rewound to catch something they missed or to make sure they did not miss something.

Discussion

Our findings suggest that viewing a group discussion through videotape rather than directly does not affect ob-
Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hit rate</td>
<td>.71</td>
<td>.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. False-alarm rate</td>
<td>.19</td>
<td>.08</td>
<td>.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Pr</td>
<td>.31</td>
<td>.13</td>
<td>.80</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Br</td>
<td>.41</td>
<td>.15</td>
<td>.77</td>
<td>.76</td>
<td>.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. No. of good observations</td>
<td>20.29</td>
<td>7.55</td>
<td>.19</td>
<td>.03</td>
<td>.19</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Elevation</td>
<td>.33</td>
<td>.27</td>
<td>.02</td>
<td>.07</td>
<td>.02</td>
<td>.09</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Differential elevation</td>
<td>.35</td>
<td>.26</td>
<td>.19</td>
<td>.18</td>
<td>.28</td>
<td>.06</td>
<td>.12</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Stereotype accuracy</td>
<td>.59</td>
<td>.24</td>
<td>.05</td>
<td>.00</td>
<td>.05</td>
<td>.03</td>
<td>.00</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Differential accuracy</td>
<td>.76</td>
<td>.26</td>
<td>.06</td>
<td>.00</td>
<td>.05</td>
<td>.03</td>
<td>.02</td>
<td>.11</td>
<td>.07</td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. No. of pauses</td>
<td>.43</td>
<td>.74</td>
<td>.08</td>
<td>.12</td>
<td>.14</td>
<td>.01</td>
<td>.28</td>
<td>.14</td>
<td>.25</td>
<td>.08</td>
<td>.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. No. of rewinds</td>
<td>1.98</td>
<td>2.90</td>
<td>.03</td>
<td>.14</td>
<td>.11</td>
<td>.09</td>
<td>.11</td>
<td>.29</td>
<td>.24</td>
<td>.10</td>
<td>.17</td>
<td>.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Helpfulness of rewinding</td>
<td>2.50</td>
<td>1.38</td>
<td>.10</td>
<td>.05</td>
<td>.06</td>
<td>.06</td>
<td>.17</td>
<td>.07</td>
<td>.29</td>
<td>.15</td>
<td>.24</td>
<td>.48</td>
<td>.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Helpfulness of pausing</td>
<td>3.74</td>
<td>1.34</td>
<td>.05</td>
<td>.12</td>
<td>.12</td>
<td>.09</td>
<td>.07</td>
<td>.19</td>
<td>.25</td>
<td>.05</td>
<td>.20</td>
<td>.49</td>
<td>.32</td>
<td>.59</td>
<td></td>
</tr>
</tbody>
</table>

Note. For Variables 1–9, N = 179. For Variables 10–13, N = 46. Pr = hit rate – false-alarm rate; Br = tendency to rate behaviors as present. *p < .05.

Observation or evaluation accuracy. Controlled observation appeared to have some effects on observation accuracy; however, these gains were not consistent or large. These results suggest that videotaping is not worthwhile if the purpose is to enhance accuracy in rating and observation. Given the costs associated with producing quality videotapes, these findings suggest that advocating their use in applied settings, under the assumption that rating quality will be improved, is unwarranted.

Controlled observation did lead to greater quality in recorded observations. This finding suggests that pausing and rewinding the videotape enabled participants to improve the specificity of what was recorded. Interestingly, participants in all conditions did a poor job of recording observations for the first target individual; the scripted behavior for this target included poor communication skills. The efficacy of controlled observation in improving rater behaviors may depend on raters’ understanding of dimensions to be assessed and the types of behaviors displayed by rates. Controlled viewing also led to improved overall ability to distinguish between present and absent behaviors. The usefulness of controlled observation in improving rater accuracy is likely to depend on the amount of noise and rater observational goals.

Post hoc analyses found differences for some of the observational-accuracy measures between videotaping of the same discussion and one of the targets. These findings indicate that the repetition of the discussion, although essentially the same in terms of verbal exchanges because of the script, led to differences in nonverbal behavior, intonations, and so forth. As there were no interactions between discussion version and condition, we could not attribute this finding to differences in the quality of the videotaping; behaviors displayed appeared to be different within the live condition as well. This finding does not limit the tests of hypotheses; rather, it highlights that participants were attending to target behaviors. However, it does suggest that further replication of a comparison of

Table 2

| Accuracy measure            | Direct |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----------------------------|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| hit rate                    | .71    | .13 | .69 | .13 | .72 | .10 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| False-alarm rate            | .19    | .08 | .21 | .08 | .18 | .07 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Br                          | .41    | .16 | .42 | .15 | .40 | .13 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Good observations           | 18.57  | 7.33| 20.13| 7.24| 22.46| 7.93|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Elevation                   | .27    | .25 | .38 | .25 | .32 | .28 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Differential elevation      | .33    | .27 | .35 | .26 | .38 | .27 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Stereotype accuracy         | .54    | .20 | .63 | .27 | .55 | .21 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Differential accuracy       | .74    | .26 | .78 | .27 | .74 | .23 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

Note. Within rows, means with different subscripts differ significantly (p < .05). Br = tendency to rate behaviors as present; Pr = hit rate – false-alarm rate.
viewing conditions with multiple versions and varied targets is warranted.

Differences might not have been found between the modes of observation because the cognitive demands placed on raters by this particular rating situation were similar. In all three viewing conditions, the rater was a passive recipient of both visual and auditory information; the rater did not interact with the ratee. In addition, the group discussion presented raters with a lot of noise (i.e., the behaviors of nontarget people) in all observation situations. Differences among observation modes for assessment center exercises involving active assessor participation, such as role-plays or questions following presentations, and with only one assesse should be examined.

There are several limitations of the study. The laboratory task may differ from an actual assessment center situation in assessor accountability and in length. Hauenstein (1992) noted that greater accountability is likely to make raters more vigilant during processing (e.g., encoding, documenting, and storing more incidents), although this does not necessarily lead to greater accuracy. The advantages of controlled observation may come into play when an assessor is fatigued, as in a daylong assessment center. The short time span of the present exercise might have allowed for assessors to remain at the same level of alertness in observation in all conditions. Research that more closely simulates the duration and the demands of a typical assessment center may be more likely to demonstrate that videotaping provides for enhanced rating accuracy. Future studies should also examine how a counterbalanced order of rating and checklist measures affects results.

Self-development through viewing videotapes and enhancement of assessor credibility are advantages that may make videotaping worthwhile, whatever the effects on rating behavior. Thornton (1992) recommended not videotaping because it might increase stress or cause participants to “play act” for the camera. However, Wiemann (1981) found that observation technology (i.e., videotaping) was not necessarily reactive (i.e., causing a change in behavior). If videotaping does not have any adverse effects, the practical advantages associated with its use (reducing the number of assessors, etc.) may make it a viable alternative to direct observation.

Videotapes are widely used in rating research (e.g., Singer & Sewell, 1989; Smith, Reilly, & Buda, 1988) and are considered an acceptable alternative to direct observation, without the criticisms directed toward paper people. The results of this study suggest that direct and indirect observational conditions may be comparable in terms of rating accuracy. Given the use of videotape in applied settings like assessment centers, further research on how controlled observation affects criteria other than rating accuracy may prove practically useful.

References


Shuster, R. (1992, September 23). No great clamor for the return of instant replay. USA Today, p. 9C.

Received August 8, 1994
Revision received May 23, 1995
Accepted May 26, 1995

New Editor Appointed

The Publications and Communications Board of the American Psychological Association announces the appointment of Kevin R. Murphy, PhD, as editor of the Journal of Applied Psychology for a six-year term beginning in 1997.

As of March 1, 1996, submit manuscripts to Kevin R. Murphy, PhD, Department of Psychology, Colorado State University, Fort Collins, CO 80523-8786.