Effects of Checklists and Feedback on Interviewer Documentation Errors

Ann Lynn Linklater
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses
Part of the Industrial and Organizational Psychology Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/833

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
EFFECTS OF CHECKLISTS AND FEEDBACK ON INTERVIEWER DOCUMENTATION ERRORS

by

Ann Lynn Linklater

A Thesis Submitted to the Faculty of The Graduate College in partial fulfillment of the requirements for the Degree of Master of Arts Department of Psychology

Western Michigan University Kalamazoo, Michigan August 1993
EFFECTS OF CHECKLISTS AND FEEDBACK ON INTERVIEWER DOCUMENTATION ERRORS

Ann Lynn Linklater, M.A.
Western Michigan University, 1993

A multiple baseline design was used to compare the effectiveness of checklists and feedback to improve marketing research telephone interview documentation. Three measures based on the combined errors of four interviewers were obtained: (1) errors per completed interview, (2) errors per interview with errors, and (3) most frequently occurring error in each of three error groups. The effects of three additional factors believed to affect interviewer performance were also investigated: (1) marketing research project, (2) complexity of the interview, and (3) lead worker responsible for editing completed interviews.

Only slight, mostly positive changes were visually evident across the dependent measures. Feedback effects were not obtained on any measure. The inferential test results showed that: (1) the project had a statistically significant effect on all measures, (2) the interview complexity had a statistically significant effect on four of the nine dependent measures, and (3) the editor was not a significant (p>.05) factor in any analysis.
ACKNOWLEDGMENTS

The completion of this thesis marks the accomplishment of a long elusive goal that would not have been possible without Brad Huitema, Dale Brethower, Alyce Dickinson, and Bill Redmon. I am deeply touched by and grateful for their investment in me and assistance with this thesis. Thank you.

Thanks also to my best friend and colleague—for everything. I especially appreciate Carol's unwavering faith in my competence, even when I doubt myself.

Finally, I dedicate this thesis to the memory of my friend, Miss Camilla Wood, for her financial and personal support in my early years of graduate school.

Ann Lynn Linklater
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700  800/521-0600

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Effects of checklists and feedback on interviewer documentation errors

Linklater, Ann Lynn, M.A.
Western Michigan University, 1993
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................. ii

LIST OF TABLES ................................................................. vi

LIST OF FIGURES ............................................................... vii

CHAPTER

I. INTRODUCTION ......................................................... 1

Performance Feedback .................................................... 1

Self-Monitoring .............................................................. 3

Productivity Measures ................................................... 4

Quality Measures .......................................................... 5

Functional Effects of Feedback ....................................... 7

Evaluation of Feedback Effects ....................................... 9

Purpose of the Study ..................................................... 9

II. METHOD ............................................................... 10

Setting ................................................................. 10

Subjects ................................................................. 11

Informed Consent ...................................................... 13

Organizational Context and Experimental Design ........ 14

Dependent Variables .................................................. 15

Predictor Variables ................................................... 17

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table of Contents—Continued

CHAPTER

Procedure ......................................................... 17
Baseline Phase .............................................. 18
Checklist Phase .............................................. 18
Feedback Phase .............................................. 19
Inter-Rater Reliability ................................. 20
Statistical Analysis ......................................... 21
Social Validity ............................................... 21

III. RESULTS ....................................................... 22
Checklist Versus Feedback ............................. 22
Average Errors Per Completed Interview .......... 22
Average Errors Per Interview With Errors ........ 22
Most Frequently Occurring Error Per Group ...... 25
Predictor Variables ................................. 25
Average Errors Per Completed Interview .......... 25
Average Errors Per Interview With Errors ........ 28
Most Frequently Occurring Error Per Group ...... 30
Social Validation ......................................... 30

IV. DISCUSSION .................................................. 32
Suggestions for Future Research .................... 37

iv

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table of Contents---Continued

APPENDICES

A. Human Subjects Institutional Review Board Approval ....................... 38
B. Error Types by Group Checklist ........................................................... 40
C. Feedback Survey Results ................................................................. 42

BIBLIOGRAPHY ....................................................................................... 45
LIST OF TABLES

1. Error Rate Means for Experimental Conditions by Error Group ............ 24

2. Coefficient of Determination and Probability Value Associated With Three Factors ........................................... 29

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF FIGURES

1. Average Errors Per Completed Interview ...................................................... 23
2. Average Errors Per Interview With Errors .................................................... 26
3. Highest Frequency Error Type in Each Error Group ................................. 27
CHAPTER I

INTRODUCTION

At the most basic theoretical level, all consequences of behavior that have some effect on the environment may provide feedback to the performer (Skinner, 1953). Skinner goes on to say that the importance of feedback is clear because organisms must be stimulated by the consequences of their own behavior if learning is to occur. Feedback may be unconditioned (e.g., sensory stimulation) or conditioned as when one's actions generate approval or affection from another. Feedback from the environment is an important part of learning to survive in the world. Contrived feedback to improve performance has been extensively studied in applied settings such as education (Leauby & Atkinson, 1989; Matthews, 1985; Whinnery & Fuchs, 1992), healthcare (Alavosius & Sulzer-Azaroff, 1989; Elixhauser, Eisen, Romeis, & Homan, 1990; Miller, 1991; Peper & Tibbetts, 1992; Rosen, Lockhart, Gants, & Westergaard, 1991), and in business and industry (Fox & Sulzer-Azaroff, 1989; Goltz, Citera, Jensen, Favero, & Komaki, 1983; Newby & Robinson, 1983).

Performance Feedback

Performance feedback is generally defined as information on the quality or quantity of an employee's past performance. Such feedback has been identified as one
of the most frequently reported primary interventions or performance management system components implemented in business organizations. The nature and effects of performance feedback interventions have been extensively reviewed (Balcazar, Hopkins & Suarez, 1986; Ford, 1980; Malott, Shimamune, & Malott, 1992; Prue & Fairbank, 1981). Although the term "feedback" is used to describe a wide variety of interventions, the significant dimensions identified by Ford and Prue and Fairbank in their reviews include: audience (public or private feedback); mechanism (verbal, written, mechanical, or self-recorded); content (group/individual data, summary or per response sample); source (self, supervisor, expert or other generated); and schedule (daily, weekly, other).

Balcazar et al. (1986) concluded that weekly feedback is as effective as daily feedback. Supervisor generated feedback was an extensively utilized source of feedback and produced the largest proportion of positive results. Relatively few of the studies reviewed involved self-generated feedback (14 compared to 54 involving supervisor generated feedback); this type of feedback yielded less consistent results. Vocal and written modes of feedback were found to be almost equally effective although graphs were the most common mode and most consistently effective. The authors' main conclusion was that feedback will lead to improved performance only if it is differentially related to other consequences such as praise, money, or other rewards. They also described the ideal feedback system as one providing graphic data, presented weekly, with performance differentially tied to tangible rewards.
Self-Monitoring

Self-generated feedback is the product of self-monitoring. Numerous studies have shown some improvement in performance through self-monitoring which typically requires self-observation and self-recording, often on a simple checklist (Brown, Malott, Dillon & Keeps, 1980; Forbes, 1982). In general, self-monitoring has been proven to be an effective tool for improving performance. In addition to the performance improvements noted, several researchers have concluded that self-monitoring positively impacts employee satisfaction (e.g., Ivancevich & McMahon, 1982; Komaki, Collins, & Penn, 1982).

Forbes' (1982) study compared the effectiveness of presenting employees with a job description that included measures and standards for the critical accomplishments of a job (often called a Job Model) with self-recording and supervisor feedback in a large residential group home. The job description was introduced first in a multiple baseline across subjects design followed by self-recording on the designated staff accomplishments. Substantial improvements were found with the introduction of the Job Model with further improvements occurring with the introduction of self-recording and feedback.

Brown et al. (1980) employed a multiple baseline with reversal design to compare self-recording on a simple checklist with training to improve retail workers' customer service behaviors. Although the checklist was limited to customer services measures, data on selling, stockwork and miscellaneous tasks were also collected.
Self-recording was found to have a significant effect on the sales staff's customer service performance but did not affect the number of sales and had little effect on overall performance. Krigsman and O'Brien (1987) found that self-monitoring and self-monitoring with quality circle feedback were both effective in improving metal clip waste but had no impact on attendance and lost work time.

Hawkins, Burgio, Langford and Engel (1992) found an increase in self-monitored prompting when written feedback was added to an existing staff management system designed to ensure staff compliance with a prompted voiding program in a nursing home. The existing system included self-recording by the nursing staff with occasional supervisory monitoring and verbal feedback. This system was considered effective prior to the case study, yet the authors were convinced that written evaluative feedback would produce even better performance.

Productivity Measures

Feedback systems have been implemented in nearly every segment of the business/organizational field. The most frequently reported productivity measures in feedback interventions are related to the quantity of work produced. For example, the number of applications and credit evaluations processed by a university admissions office increased substantially with the implementation of a daily adjusted goal setting and feedback system (Wilk & Redmon, 1990).

Other performance related areas targeted for feedback interventions include
compliance with established procedures in a mental health organization (e.g., Calpin, Edelstein & Redmon, 1988), absenteeism (Krigsman & O'Brien, 1987), operator setup time in a manufacturing setting (e.g., Wittkopp, Rowan, & Poling, 1990), and safety practices of workers in a poultry processing plant (Komaki et al., 1982).

Quality Measures

Until the Total Quality Management (TQM) movement began in recent years, relatively few interventions employed feedback to improve the quality of products or accomplishments. While not the focus of the intervention, in many feedback studies a quality measure is implicit in the dependent variable (e.g., when the quantity or rate of performance must meet quality standards). For example, Henry and Redmon (1990) demonstrated an increase the percentage of correctly completed tasks required for an SPC program while retaining the quality of parts produced. In many quality improvement programs, feedback is usually only one of many components of a system that may include goal setting, participative management, and performance incentives.

Quality is usually evaluated through inspection of parts, either throughout the process or when the product is finished; yet, human inspection tends to result in low levels of defect or error detection (Harris & Chaney, 1969). Factors contributing to the accuracy of visual inspection have been identified as: supervision or form of feedback (Drury & Addison, 1973), the probability of an error or defect occurring (Fortune, 1979), and the nature and complexity of the task (Harris & Chaney, 1969).
Rigby and Swain (1975) stated that the greater the number of possible errors, the greater the likelihood that many will be overlooked. They also suggest that this is especially important when workers must do their own error checking (e.g., editing a report, scanning the finished product for defects).

Using a computer simulated inspection task, Mason and Redmon (1992) studied (A) pacing type (self-paced versus machine-paced) and (B) lag in presentation of the feedback (immediate versus delayed) on the accuracy of identifying errors in a sample "product." They found the following rank order on accuracy of four conditions: (1) self-paced, immediate feedback, (2) self-paced delayed feedback, (3) machine-paced immediate feedback, and (4) machine-paced delayed feedback. Hence, accuracy was directly related to the source of pacing and immediacy of feedback. This study took place in a laboratory setting in which the subjects had few if any of the distractions found in quality inspection in the actual workplace, but suggests areas for future research on the feedback parameters related to quality inspection.

Frederikson, Richter, Johnson, and Solomen (1981) used a reversal design with two groups of employees responsible for recordkeeping in a university psychology clinic. Feedback was provided and then withdrawn on two of the four types of recordkeeping errors (completeness, status, format, and signatures) with each group. Errors decreased with the introduction of feedback on the targeted errors without any systematic effect on the non-targeted errors. In other words, feedback affected only the quality components of the task for which it was given.
There is little agreement but much discussion in the Organizational Behavior Management (OBM) field about precisely how feedback works, beyond its motivational function. Early analyses of the functional effects of feedback concluded that feedback may serve as a reinforcer or discriminative stimulus (Balcazar et al., 1986; Peterson, 1982), and possibly as an establishing operation (Duncan and Bruwelheide, 1986), as defined by Michael (1982). Daniels (1989) states that feedback only provides an opportunity to improve performance, but if consistently associated with positive reinforcement, feedback assumes the properties of conditioned reinforcement. However, given the lengthy temporal delay between behavior and feedback that is typical in organizational settings, many argue that feedback typically does not function as a simple reinforcer (e.g., Agnew & Redmon, 1992; Duncan & Bruwelheide, 1986; Malott et al., 1992).

Peterson (1982) suggested that infrequent or temporally delayed feedback may function as rule governed behavior (RGB) rather than contingency shaped behavior. For example, if a supervisor gave an employee a feedback graph displaying the quantity of widgets she produced in the preceding month, without any verbal instructions or comments regarding the feedback, the employee might state a rule such as "If I make more widgets, my supervisor will nominate me for a promotion." Agnew and Redmon (1992) suggest that this rule could alter the function of stimuli associated with making more widgets (e.g., work to be completed could become a discriminative
stimulus affecting the amount of widgets produced and knowledge of the widgets produced could become a reinforcing stimulus). Once the rule is stated, those stimuli might maintain the behavior of producing more widgets.

Agnew and Redmon (1992) and Malott et al. (1992) also agree that performance feedback may generate rule governed behavior, based on the receiver's previous experience within an organization. Agnew and Redmon (1992) further suggest that it is best for management to specify the rules for performance feedback rather than letting employees generate their own.

Locke, Cartledge and Koeppel (1968) stated that performance feedback always implicitly implies a performance standard and therefore affects motivation to the degree the recipient values improving his or her own performance, especially in the absence of a standard. In other words, the effectiveness of any performance data depends on whether the employee considers it significant and sets goals around it. In later studies, Locke and Latham (1990) stated that feedback without goals has little effect on performance and vice versa. That is, without a goal or standard, Locke and Latham believe that people don't see feedback as significant and will not take action in response. They also suggest that the need for goals is especially obvious when employees are confronted with multiple types of feedback.

At the very least, feedback probably has multiple functions which may not be readily obvious in the organizational environment. Given what is and is not known about how best to implement feedback in a specific organization to obtain optimal
results for the cost, it remains important for managers to evaluate proposed feedback systems in light of the organization's needs, environment, and culture.

Evaluation of Feedback Effects

Research designs would seem to be the ideal method for evaluating decisions in the regular work environment. Unfortunately, the typical organization carries out less rigorous (or no) evaluation efforts when implementing performance improvement interventions (Geroy & Wright, 1988). The authors go on to say that the reasons for this lack of rigor are usually related to time and cost constraints as well as a tendency of decision makers to rely on their "gut feelings" about the outcome of certain interventions. These factors are worthy of management consideration, given that the costs of evaluation must be borne out for some time without any tangible payback. However, the costs of evaluation may prove to be worthwhile if the results show that the least costly intervention is equally or more effective than another intervention.

Purpose of the Study

The major purpose of the present study was to evaluate the effectiveness of performance feedback and checklist procedures in the reduction of errors in an interview documentation task in a marketing research department. A second purpose was to evaluate the relationship of three potential predictors of errors: project, complexity of the interview, and editor.
CHAPTER II

METHOD

Setting

This study took place in the Marketing Research Department of a large, international direct marketing company. The company manufactures the majority of its own products that are sold by independent distributors. The Marketing Research department is responsible for designing and conducting telephone interviews with the company's independent distributors, and less frequently, retail consumers and company staff. Any department may request market research on topics ranging from service evaluation to changes in packaging or new products. Most marketing research projects are completed within two days but some may require as many as nine working days. Occasionally, one project may be started while a second is still being conducted. Projects also differ on the degree of complexity of the interview form and the number of interviews to be completed.

A marketing research project is completed in several steps: (a) The interview survey is designed by specialists, (b) telephone interviewers talk to respondents on the telephone and document their responses on an interview form, (c) Project Coordinators then edit the completed interview form to ensure complete and accurate data have been obtained, (d) data coders next analyze the data, and (e) a report is sent.
to the department that requested the interview information.

The Marketing Research telephone interviewers must follow a rigid procedure to avoid biasing respondents' answers or contaminating the data obtained. The interviewers are also required to document the responses clearly and accurately for data coding and analysis. Interviews with incomplete and/or unclear responses necessitate a call-back to obtain usable data.

Telephone interviewer performance data are collected by two lead workers (known as Project Coordinators). The Project Coordinators monitor actual interviews by listening in on a 10 minute portion of an interview for each interviewer working on a project day. They then provide each interviewer with a feedback summary at the end of the day. Although Project Coordinators edited each completed interview for content prior to this study, they did not formally track data or give the telephone interviewers feedback on interview documentation errors unless call-backs were required. The two Project Coordinators alternate monitoring and editing activities by project, although both may do editing on the same project if needed.

Subjects

The evaluation participants were eight telephone interviewers who worked evenings unless interviews had to be conducted during the day by respondent request or for special projects. The seven female and one male subjects ranged in age from 20 to 45 years. All telephone interviewers received training on how to conduct and
document telephone interviews upon being hired. This training ran approximately two
to three days and included interactive lecture, self-study using a workbook, a video
showing an exemplary interviewer on the telephone, and practice interviewing.
Content areas included professional interviewing techniques (e.g., introductions,
probing, handling respondent concerns) and documenting interviews (e.g., note taking,
following skip patterns, verbatim reporting). At the time of this study, three of the
eight part-time interviewers were new hires and had just completed the interviewer
training. The remaining staff had been working for several months to several years. All
telephone interviewers worked part time as needed for projects. This meant that staff
might not have work for as few as one to two days or as long as one month. One to
eight staff members worked on any given night.

This study is based on group data of only four of the telephone interviewers
described previously. These four were selected as the subjects because they worked on
the same days the most consistently. This was necessary because not all of the same
telephone interviewers worked on any given project day. It is desirable to have the
same group of performers in order to meaningfully compare performance across days.
As the data from this group are not confounded by changes in group composition,
comparisons across conditions are more easily interpreted.

Removing all data from project days on which fewer than three interviewers
were working reduced the data set by 15 work days (48 instead of 63 data points in
the general evaluation). After deleting those 15 data points, there were 14 project days
on which only one of the four interviewers did not work. Because each of the four interviewers was absent at least once on different days during each of the three phases, an estimate of the number of completed interviews and errors for the one absent interviewer was calculated for an additional 14 project work days. This was done to ensure that each phase had as many data points as possible for the statistical analysis. This approach also seemed reasonable since data were available for the other three interviewers. The estimates were based on the average of the single interviewer's performance data on the work day immediately before and after an absence.

Informed Consent

All interviewers were routinely monitored on the telephone prior to the interview documentation feedback system. Staff members were hired with full knowledge that they would be monitored as part of the monitoring and feedback system. No interviewer was given the option of not participating in the evaluation. No interviewer was penalized for his or her performance during the evaluation. The Western Michigan University HSIRB determined that it would not be necessary nor appropriate to request a signed consent form to use group data for the purpose of this thesis. (The HSIRB letter of approval can be found in Appendix A.) All telephone interviewers were asked to complete an anonymous survey regarding the feedback system approximately two months after the evaluation ended. They were given the opportunity to decline to participate by not returning the completed survey.
Organizational Context and Experimental Design

The Marketing Research Department had a telephone interview monitoring and feedback system in place prior to initiation of the current study. A performance management consultant was initially hired to design a pay for performance system for the Department. The consultant determined that it would first be necessary improve the performance measurement system which had been operating for approximately one year. The management also requested that the consultant add an interview documentation monitoring and feedback component to the existing system. The consultant worked with the Project Coordinators to formalize the editing process while designing the interview documentation component. The Marketing Research Department management was involved in all stages of the consultant's work in improving the original performance monitoring system, including the design of the evaluation.

The focus of this study was entirely on the written documentation of the interviews. A multiple baseline across interviewer behaviors (error types) research design was used to compare interviewer documentation performance under three conditions: baseline, checklist, and feedback. The multiple baseline design was selected because (a) it provides good internal validity (e.g., Komaki, 1977), and (b) it is practical. The checklist versus feedback comparison was of interest to determine whether the Project Coordinators' time would be well spent providing feedback.
Dependent Variables

Fourteen types of mistakes that telephone interviewers can make as they write down a respondent's answers to survey questions were formally identified by the consultant in her work with the Project Coordinators. (The complete list of error types by group can be found in Appendix B.) Some errors resulted from poor writing (e.g., writing not legible, punctuation omitted, or abbreviation errors) that affected the data coder's ability to correctly interpret a response. Other errors could occur when the interviewer failed to follow the skip pattern instructions for each question. For example, if the interviewer did not correctly follow the instruction to skip to question 3, or continue with the next question depending on the respondent's answer to question 1, a "skip pattern" error occurred. These error types were split into three Error Groups in order to implement the multiple baseline design.

The three Error Groups were intended to be balanced across frequency of the different errors. The frequency of each type of error tracked during the baseline phase (prior to implementing the first checklist) was calculated to ensure that each Error Group contained some high, medium and low frequency errors. Each Error Group was also established to include at least one serious error to ensure that no one grouping had all serious errors. Errors were defined as serious if they necessitated a call-back or otherwise invalidated the interview data (errors A, B, D, P, T, and V in Appendix B). Management rated each type of error as to its importance based on the seriousness of the problem resulting from the error.
Three different measures based on the items in each Error Group were analyzed: (1) the average errors per completed interview, (2), the average errors per interview with errors, and (3) the most frequently occurring error in each Error Group.

Although each of these measures was based on a single group of interviewers (n = 4), it was useful to calculate the average error per completed interview because there were differences in the total number of interviews each interviewer completed on a given day. One extremely high observation in Error Group 3 was believed to be of questionable validity due to a bias in the editing criteria. Therefore, the average of the Group 3 total errors across the 48 observations was substituted for the extreme value.

The average errors per interview with errors was investigated because each interview form provides multiple opportunities for any number of different errors to occur. Thus, even when the total number of errors did not significantly decrease, if the total number of completed interviews declined, the average number of group errors associated with each interview with errors would show higher rates than the average calculated for all completed interviews. The adjusted observation value used for the previous measure was again substituted for the questionable observation.

Each Error Group contained one high frequency error that could account for much of the variability within an Error Group. Thus, it makes sense to look at the contribution of the most frequently occurring error on the average error rate for each group. The average number of "K" errors produced by all four interviewers was substituted for the questionable observation in Error Group 3.
Predictor Variables

Although the focus of the study was on the comparison of the 3 conditions (baseline, checklist, and feedback), there was also interest in evaluating the effects of additional factors believed to affect the occurrence of errors. The nature of the project and systematic differences between editors on error identification were hypothesized as potential explanations of variability in errors committed.

Projects are unique as they vary by content, purpose, structure. The "projects" factor was operationalized two ways. First, dummy variables were constructed to identify the 16 different projects; this set of variables was employed in the statistical analysis to evaluate the importance of project differences in explaining variation in interviewer errors. Second, the complexity of the interview projects was considered. Interview complexity was measured on a 3 point scale (easy, medium, and difficult) based on the average subjective rating of the two Project Coordinators, the consultant, and the author.

Each of the two Project Coordinators alternated editing and telephone monitoring by project. Each was assigned a number to identify who had done the editing on each project day. A third number was used to identify the few days when both Project Coordinators edited on the same or two different projects.

Procedure

Each experimental condition or design phase was intended to last for a
minimum of five working days. Some phases lasted longer because of extended time off between projects due to lack of projects. Also, phase changes were scheduled to avoid coinciding with the start of new projects to minimize errors associated with learning new interview patterns. The actual duration of the phases ranged from 9 to 16 project work days to accommodate those issues.

**Baseline Phase**

The baseline condition involved the collection of documentation error data on 14 behaviors prior to any intervention. All subjects had been exposed to the initial training. No changes were implemented during this phase although interviewers did receive vocal feedback from Project Coordinators on serious errors as needed.

**Checklist Phase**

The checklist condition consisted of performers receiving a checklist of the errors for each new group plus the errors from any previous checklist. For example, the first baseline checklist (Error Group 1) had four error types, the second baseline checklist had another four error types (Error Group 2) plus those from Error Group 1 for a total of eight error types. The third baseline checklist had a total of 14 errors listed (six from Error Group 3 plus the other eight). At the start of work at each checklist condition change, the supervisor gave each telephone interviewer a checklist. The interviewers also received the oral instruction to use the checklist to check each
interview before giving it to the Project Coordinator for editing. No other prompts were given after the checklist was handed out, with the exception of one extra prompt from the supervisor near the end of the evaluation. The performers were not required to complete and attach a checklist to each interview.

Feedback Phase

The feedback condition was defined as one in which the Project Coordinator attached a feedback form to each interview with documentation errors and gave it to the interviewer for review. The Project Coordinator responsible for editing attached a feedback form to each completed interview containing one or more of the errors listed on the relevant checklist. When the editor found a call-back error, the interview form was immediately returned to the interviewer with verbal instructions to correct the problem. In other words, the interviewers were informed of call-back errors regardless of which Error Group feedback condition was in effect. Only the type of error and question number were indicated on the feedback form. The interviews with feedback (other than call-backs) were returned to the performer as soon as feasible. This occurred during work, at the end of the night, the next day, or whenever the performer returned to work. The Project Coordinators did not provide evaluative statements nor discuss the written feedback with individual interviewers unless approached by the interviewer.
Inter-Rater Reliability

Telephone interviewers have to make some fairly difficult discriminations to meet the Department's standards for reliable, accurate interview data. This is also true for those responsible for monitoring the quality of the interview documentation. The consultants did not feel qualified to make those discriminations and felt the more important issue was inter-rater reliability between the editors. Therefore, the two Project Coordinators were asked to independently review one randomly selected interview. As the consultant was concerned about adding to the Project Coordinators' already heavy workload with frequent reliability checks, the inter-rater reliability trials were only conducted at six different points, beginning with the first feedback phase.

Due to differences in the interviews across projects and the number of possible errors and opportunities for errors, inter-rater reliability was calculated as the number of agreements divided by the number of agreements plus disagreements between the two Project Coordinators. An agreement was defined as consistency between the number and type of errors and a disagreement was defined as any discrepancy between the number and/or type of error. The average agreement for the six reliability values (100, 100, 93, 74, 73, and 73%) was 86%. The pooled average, calculated as the total number of agreements divided by the total number of agreements and disagreements, was 85%. The reliability data were not disseminated during the evaluation.
Statistical Analysis

Time series regression procedures were used to analyze intervention effects of checklist and feedback conditions and to evaluate relationships between predictor variables and errors. The details of the intervention model can be found in Huitema and McKean (1991, pp. 4-8). Computation was performed using Minitab Release 9.1 software running on a VAX computer.

Social Validation

Approximately two months after the evaluation ended, all Marketing Research telephone interviewers were asked to anonymously complete a seven-item survey. Each interviewer was provided a self-addressed stamped envelope to return the paper survey form to the author. The survey questions (approved by management) were designed to determine the subjects' opinions of the feedback system and their own error checking strategies. (The survey questions can be found with the results in Appendix C.)
CHAPTER III

RESULTS

The results of the primary intervention (i.e., checklists and feedback) will be presented first. These results are displayed in graphic format with condition means with the results of significance tests presented in a table. Subsequently, results on the three predictor variables will be presented.

Checklist Versus Feedback

Average Errors Per Completed Interview

Figure 1 shows the average number of errors per completed interview in each condition by Error Group across time and conditions. Note that the average error rate changed very little on any Error Group when the checklist for Error Group 1 was introduced. In both Error Groups 2 and 3, the mean error rate decreased by approximately one half unit when the feedback was introduced in Error Group 1. The means for each experimental condition are displayed in Table 1 along with the results of overall tests on differences among phase levels. None of the tests is significant ($p>.05$).

Average Errors Per Interview With Errors

Figure 2, shows that the patterns within and across the 3 Error Groups did not
Figure 1. Average Group Errors Per Completed Interview.

* Includes estimate for one of the four subjects.
* Group average value substituted for questionable observation.

Extended phase means indicated by _______ line, condition means by _______ line.
<table>
<thead>
<tr>
<th>Error Group</th>
<th>Average per completed interview</th>
<th>Average per interview with errors</th>
<th>Highest frequency error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Group 1</td>
<td>.47</td>
<td>.46</td>
<td>.36</td>
</tr>
<tr>
<td>Error Group 2</td>
<td>.23</td>
<td>.08</td>
<td>.19</td>
</tr>
<tr>
<td>Error Group 3</td>
<td>.24</td>
<td>.21</td>
<td>.19</td>
</tr>
<tr>
<td>Error Group 1 - &quot;B&quot;</td>
<td>6.12</td>
<td>5.14</td>
<td>4.92</td>
</tr>
<tr>
<td>Error Group 2 - &quot;D&quot;</td>
<td>3.31</td>
<td>.51</td>
<td>2.70</td>
</tr>
<tr>
<td>Error Group 3 - &quot;K&quot;</td>
<td>1.89</td>
<td>0.13</td>
<td>0.00</td>
</tr>
</tbody>
</table>

ns = non significant
*p < .05.
change with respect to the average errors per interview with errors. Again, Table 1 shows that none of the changes in the experimental conditions led to statistically significant differences in errors (\(p > .05\)).

**Most Frequently Occurring Error Per Group**

Each Error Group contained one high frequency error that accounted for most of the errors (75% in Group 1, 77% in Group 2, and 38% in Group 3). Figure 3 shows that the condition means fell slightly immediately when the checklist was implemented but stayed relatively high in the feedback condition for Error Groups 1 and 2. The overall test on differences among phase means was significant (\(p < .05\)) in Group 2 only (see Table 1).

While the level change across phases is not large, it is obvious in Figure 3 that a change in variance has occurred in Error Group 3. A comparison of the baseline phase variance with the variance for the other phases reveals a significant reduction in the latter using an F test (\(p < .001\)).

**Predictor Variables**

Results on the relationship between each error measure and the predictor variables of project classification, interview complexity, and editor classification are presented in this section.

**Average Errors Per Completed Interview**

It can be seen in Table 2 that the marketing research project being conducted
Figure 2. Average Group Errors Per Interview With Errors.

* Includes estimate for one of the four subjects.
* Group average value substituted for questionable observation.
Figure 3. Highest Frequency Error Type in Each Error Group.

- Includes estimate for one of the four subjects.
- Average of the four subjects' total errors substituted for questionable observation.
on any given day accounted for much of the variation in errors. The projects factor explained 63, 53, and 55% of the variation in Error Groups 1, 2 and 3 respectively. This effect is statistically significant ($p<.01$) with respect to Group 1 and ($p<.05$) for Error Groups 2 and 3.

The complexity of the interview is one possible explanation for why the project factor explains much of the variability. Notice that the interview complexity had a highly statistically significant ($p<.001$) relationship with Group 2 errors and explained 26% of the variation on this measure. While this factor explained less of the variability (4 and 11%) in Error Groups 1 and 2, respectively, the relationship was significant ($p<.01$).

Although the editor classification was expected to be an important factor, no effects were found ($p>.05$) on any Error Group measure.

**Average Errors Per Interview With Errors**

The projects factor was also highly significant ($p<.01$) in all three Error Groups. Table 2 shows that the project classification explained approximately 60% of the error variation on all Error Groups.

The complexity of the interview accounted for 21% of the variation in Group 2 errors. This effect was significant ($p<.01$). No effects were found on the other Error Groups.

Again, the editor characteristic had no significant effect on any group measure.
Table 2

Coefficient of Determination and Probability Value Associated With Three Factors

<table>
<thead>
<tr>
<th>Error rates</th>
<th>Factor</th>
<th>Project</th>
<th></th>
<th>Interview complexity</th>
<th></th>
<th>Editor</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>p</td>
<td>R²</td>
<td>p</td>
<td>R²</td>
<td>p</td>
</tr>
<tr>
<td>Average per completed interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error Group 1</td>
<td></td>
<td>.68</td>
<td>***</td>
<td>.04</td>
<td>*</td>
<td>.03</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 2</td>
<td></td>
<td>.61</td>
<td>**</td>
<td>.26</td>
<td>***</td>
<td>.02</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 3</td>
<td></td>
<td>.66</td>
<td>**</td>
<td>.11</td>
<td>*</td>
<td>.00</td>
<td>ns</td>
</tr>
<tr>
<td>Average per interview with errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error Group 1</td>
<td></td>
<td>.59</td>
<td>**</td>
<td>.00</td>
<td>ns</td>
<td>.05</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 2</td>
<td></td>
<td>.60</td>
<td>**</td>
<td>.21</td>
<td>**</td>
<td>.01</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 3</td>
<td></td>
<td>.58</td>
<td>**</td>
<td>.01</td>
<td>ns</td>
<td>.00</td>
<td>ns</td>
</tr>
<tr>
<td>Highest frequency error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error Group 1 - &quot;B&quot;</td>
<td></td>
<td>.63</td>
<td>**</td>
<td>.02</td>
<td>ns</td>
<td>.00</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 2 - &quot;D&quot;</td>
<td></td>
<td>.53</td>
<td>*</td>
<td>.07</td>
<td>ns</td>
<td>.04</td>
<td>ns</td>
</tr>
<tr>
<td>Error Group 3 - &quot;K&quot;</td>
<td></td>
<td>.55</td>
<td>*</td>
<td>.02</td>
<td>ns</td>
<td>.00</td>
<td>ns</td>
</tr>
</tbody>
</table>

ns indicates $p > .05$.  $^* p < .05$.  $^* * p < .01$.  $^* * * p < .001$.  

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
**Most Frequently Occurring Error Per Group**

The projects factor explained less of the variability in errors on this measure (63, 53, and 55% in Error Groups 1, 2, and 3 respectively) than on the average errors per completed interview or average error per interview with errors. Table 2 shows that these effects were also significant for all three Error Groups (p<.01 for Error Group 1 and p<.05 in the other Error Groups).

The complexity of the interview accounted for little of the variation in errors (2% in Error Groups 1 and 3) and (7% in Error Group 2). These percentages were not statistically significant (p>.05).

As with the other measures, no effects for the editor factor were found.

**Social Validation**

Three of the six respondents to the Feedback Survey said they felt their work was as accurate when they checked their own work as when they received feedback from the Project Coordinators. (A summary of the survey results can be found in Appendix C.) All six respondents reported that they did reread each completed interview to check spelling, completeness, and content accuracy before submitting it to the Project Coordinator. Two of the telephone interviewers reported that how well the interview survey was written always affected their ability to complete the interview, with two others saying the interview always affected their performance until they
became familiar with the interview.

Five of the six telephone interviewers indicated that they preferred receiving information about errors in person and that they liked to talk with Project Coordinators about feedback on errors. None preferred written feedback and all six respondents said that receiving feedback on good performance was important to them. As one respondent said "It's very important to know when you made a mistake but it's just as important to know when you are doing a good job."

Of the four interviewers providing comments on the feedback they received, all said they felt some of it was incorrect or too "nit-picky." One respondent noted that "...sometimes I feel we get too much feedback. When we get reports every night, I feel the value of the feedback loses some of its importance."
CHAPTER IV

DISCUSSION

The feedback system evaluated in this study did not produce a significant improvement in the quality of the interview documentation as measured by the number of errors. This finding is not unique as inconsistent results have frequently been noted in the feedback literature (e.g., Balcazar et al., 1986; Prue and Fairbank, 1981).

Although the checklist appeared to produce slight improvements, it was only statistically significant for the highest frequency error measure in Error Group 2. Presentation of the checklist was probably most effective as a job aid. In this case, the checklist condition did not fulfill the definition of self-monitoring because the interviewers were not required to self-record when checking a completed interview. However, the checklist appeared to be an effective antecedent because it formally specified the types of errors that the Project Coordinators would be attempting to identify in the editing process. Prior to this study, the telephone interviewers received instruction on how to accurately document an interview (i.e., avoid the 14 types of errors) in the training. They also received feedback on errors when the Project Coordinators notified them when a call-back was required (e.g., a response was not clear and therefore could not be coded).

The feedback as implemented did not have an effect on errors. Yet the
previous research on performance feedback provides clues for analyzing why the system as implemented was not more effective. For example, Rigby and Swain (1975) argued that employees can only respond to a limited number of errors in visual inspection. The large number of possible errors and complexity of the task may have mitigated the interviewers' ability to edit their own work effectively or to respond to the feedback provided in this study.

Despite having operationally defined and validated the 14 documentation errors prior to implementing the evaluation, the telephone interviewers questioned the reliability and validity of the feedback they received. In essence, there are three issues relevant to the feedback provided: (1) inter-rater reliability in error identification, (2) the accuracy of errors identified by the editors, and (3) the importance of information to interviewers.

The data showed that the Project Coordinators often failed to agree that a given error had occurred in the inter-rater reliability tests. Even though the overall reliability was 85%, the interviewers could contest any "error" they did not feel was justified. Since the interviewers were focusing on individual errors instead of patterns, the feedback probably did not convey information to the telephone interviewers that they (a) valued and (b) could use to improve future performance. Locke et al. (1968) suggested that performance feedback must do both to be effective.

Accuracy is defined as the extent to which identified errors fit the criteria of the definition. Despite having operationalized the 14 errors, only two of the six telephone
interviewers responding to the Feedback Survey said they felt confident that the feedback they had received was usually accurate. If the editors’ accuracy was indeed low, the interviewers would not be able to use the feedback to improve their performance. As none of the Feedback Survey respondents felt that the feedback they received identified serious errors, the minimal effects of the checklist and feedback conditions may have been due to the quality and accuracy of the feedback received rather than the information provided.

Telephone interviewers received only negative feedback on their interview documentation errors and tended to feel that much of the feedback focused on trivial errors or was inaccurate. As no standards were set for performance and no rules were established as to how the interviewers were expected to perform, it is likely that any rules they generated were stated negatively (i.e., do better to avoid negative consequences), a situation which Agnew and Redmon (1992) recommend avoiding. As the feedback sheet was attached to each interview form that contained errors, it is likely that the telephone interviewers were not able to gauge the extent of their errors each work day or to specify any "rule" related to changing their performance over time. This seems especially probable when the interviewers had several days off between project work days. Feedback presented as summary data in graphic format would have allowed the interviewers to see trends in their performance.

The feedback system as implemented did not provide any back up consequences to the knowledge of results provided. There is a consensus in the field
that better and more long lasting effects are achieved when other rewards or incentives (e.g., money, job enhancement, etc.) are part of the performance management system (Balcazar et al, 1986; Daniels, 1989). In effect, the feedback as implemented did not provided differential consequences which Balcazar et al. concluded are necessary for feedback to be effective. Although the telephone interviewers appeared to appreciate knowing how well they were doing, the interview documentation feedback may not have been strong enough to motivate improved performance without additional management support.

Fedor and Buckley (1987) suggested that frequent feedback may not facilitate better performance and is possibly dysfunctional. Considering the amount of feedback provided to the interviewers through the original telephone monitoring system, the addition of the documentation feedback component may have made this environment too "feedback dense" to be effective. It is possible that because feedback was already being provided on the telephone monitoring component each work day, it is not realistic to expect to see the dramatic improvements found in other feedback research (cf. Wilk & Redmon, 1990). In other words, substantial improvements would be more evident in organizations or areas in which feedback was not currently supplied (cf. Hawkins et al., 1992).

The most important factors affecting error rates were found to be inherent in the marketing research project, which varies by interview design, type of project (i.e., purpose and content), and the Project Coordinator responsible for editing.
The interview complexity explained a significant percentage of the variability on errors. Despite having been thought to be an important factor, the editor classification was not found to be significant in any case. But these results are not surprising in light of the research on quality inspection (cf. Drury & Addison, 1973) and the complexity of the tasks involved in telephone interviewing.

Harris & Chaney (1969) identified task complexity as a factor influencing the accuracy of human quality inspection. In this study, the performers stated that the marketing research interview survey made it difficult to conduct the interview accurately despite feeling confident that they usually knew what was necessary to complete an interview correctly. It is not clear whether the checklist and/or feedback conditions influenced the interviewers' accuracy while conducting the interview or correcting errors when checking over the completed form. With regard to the latter, the checklist and feedback results are consistent with the research on quality inspection which shows that inspection accuracy is a function of the probability of errors occurring and task complexity.

Despite the general lack of significant checklist and feedback results obtained in this study, the evaluation design and use of inferential statistics did provide a framework for analyzing the additional factors contributing to the quality of interviewer documentation and for identifying potential improvements to the feedback system. This study supports the need to rigorously evaluate any feedback system in terms of the environmental and cultural variables of the organizational environment to
ensure the most cost effective performance management intervention. The results of the current evaluation support previous research suggesting that feedback systems should (a) provide positive as well as negative feedback, (b) focus on specific, critical measures, and (c) specify the contingencies for the performers (i.e., tell employees how their performance will be evaluated as well as the expected standards).

Suggestions for Future Research

Despite the lack of significant feedback and checklist results, this study provides some direction for future research on the design and evaluation of feedback systems in complex organizations. The relationship among the complexity of the task and project design should be studied in order to determine whether feedback is the most cost-effective intervention. That is, it may be more effective to modify the complexity of the task or aspects of the project than to change the feedback system. Only after an analysis of a performance problem in the total context of the environment should feedback be considered the best solution. Although the checklist was an antecedent for correcting errors committed during the interview documentation, it is not clear what behavior the feedback was associated with. Therefore, it would also be useful to investigate how feedback differentially affects interviewer documentation and error checking tasks. Additional research is also needed to identify the optimal amount of information an employee can reasonably be expected to respond to, especially when multiple sources of feedback are provided.
Appendix A

Human Subjects Institutional Review
Board Approval
To: Ann Linklater
From: Mary Anne Bunda, Chair
Re: HSIRB Project Number: 91-03-20

Date: March 20, 1991

This letter will serve as confirmation that your research protocol, "Comparing the Effectiveness of Self-Monitoring (checklist) with Mediated Feedback to Evaluate the Performance System Components," has been approved under the exempt category of review by the HSIRB. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the approval application.

You must seek reapproval for any changes in this design. You must also seek reapproval if the project extends beyond the termination date.

The Board wishes you success in the pursuit of your research goals.

xc: Dale Brethower, Psychology

Approval Termination: March 20, 1992
Appendix B

Error Types by Group Checklist
EDITING CHECKLIST

- (A) Skip pattern error
- (C) Writing not legible
- (K) #'s not in red
- (T) Closed end omitted
- (V) Open end omitted
- (X) ADA# omitted
- (F) Punctuation omitted
- (E) Circled # location not correct or omitted
- (D) w/e (what else) omitted
- (P) Response not clear/ Doesn't answer ques.
- (J) Abbreviation errors
- (G) Heading blanks not completed
- (B) Specific probe omitted
- (H) Spelling not comprehensible
Appendix C

Feedback Survey Results
FEEDBACK SURVEY RESULTS

1. Do you think your work is as accurate when you check your own work as when you get feedback from a Project Coordinator? Please explain why.
   
   _3_ Yes
   _1_ No
   _2_ Sometimes

   1. There are times the PC checks your work and doesn't understand your answers, but when she comes to you and you explain it, another conclusion results. (sometimes)
   2. Sometimes you miss something that a project coordinator may catch. (no)
   3. When editing a survey I may miss something I should have probed. The PC will bring it to my attention. (sometimes)
   4. I feel I am pretty accurate. (yes)
   5. I try to find my own mistakes so the PC doesn't have to give me any feedback if possible or only that I did good. We have a special edit sheet that we follow and it helps me. (yes)

2. When is it most useful for you to get feedback on your work from a Project Coordinator?

   _1_ After completing each interview
   _1_ When problems occur
   _4_ At the end of a work shift
   _0_ At regularly scheduled performance reviews every 6 months
   _0_ The next day

3. How often do you check your own work before submitting a completed interview to a Project Coordinator?

   _6_ Every interview
   _0_ Never
   _0_ Once per work shift
   _0_ Occasionally or no particular pattern

4. What kinds of things do you do to check your own work?

   _6_ Reread questions and answers for:
      
      _3_ Clarity
      _3_ Complete responses
      _1_ Skip patterns followed correctly
      _2_ All questions answered
      _1_ Legibility
      _2_ Spelling
5. Please check each of the statements below that describes your feelings:

- 6. Getting feedback when I perform well is important to me
- 5. I prefer having some one tell me in person when there are errors in my work
- 5. Knowing the number of errors in my work is important to me
- 5. I like to talk about my work with the Project Coordinator when I get feedback on errors
- 5. I feel that I usually know what I have to do to complete an interview correctly
- 2. I feel confident that the feedback I receive is usually accurate
- 0. I feel that the feedback I get usually identifies serious errors
- 0. I prefer written rather than verbal feedback when errors occur

6. Please think back over the last 5 or 6 projects you have worked on. How often did way the interview was written affect how well you were able to complete the interview?

- 2. Once or twice
- 2. Always, until I became familiar with the interview
- 2. Every interview
- 0. Never
- 0. Most of the time

7. Please write any comments or suggestions regarding the above items. (n = 4)

- 2. Sometimes the feedback is trivial or petty.
- 1. Projects can help or hinder doing the interview.
- 2. Feedback is mostly on what is wrong / little positive feedback.
- 1. Problems occur when you and the PC don't agree that something is wrong.
- 1. Most of the feedback I agree with or see why it was given.
BIBLIOGRAPHY


