DATE DUE	DATE DUE	DATE DUE
JAN 3 0 2001 062701		

DEVELOPMENT OF AN INSTRUMENT TO ASSESS THE IMPLEMENTATION FIDELITY OF PHYSICAL EDUCATION LESSONS

By

Hasan Talal Al-Tawil

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Kinesiology

1998

ABSTRACT

DEVELOPMENT OF AN INSTRUMENT TO ASSESS THE IMPLEMENTATION FIDELITY OF PHYSICAL EDUCATION LESSONS

By

Hasan Al-Tawil

The purpose of this study was to develop an instrument to measure the degree to which physical education teachers, who agree to do so, implement K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project (MI-EPEC) as intended by lesson developers. Seven steps were used to develop the instrument.

Step 1: Determine the purpose and objectives of the instrument

Step 2: Construct a table of test specifications

Step 3: Review and revise an item pool

Step 4: Prepare instructions for using the instrument

Step 5: Use and refine the instrument

Step 6: Establish the content validity of the instrument

Step 7: Establish the reliability of the instrument

The data source for the first six steps was the MI-EPEC implementation team (MIT). Members of this team included the project's curriculum and instruction specialist, evaluation specialist, and the project co-chair who fulfilled both curriculum and evaluation functions. For step seven, a student rating team (SRT) consisting of four students enrolled at the University of Michigan, was used. All data were collected between April 30 and July 2, 1998. A brief overview of the results obtained for each step follows.

For step 1, one statement of purpose and three objectives were created.

In step 2, data were obtained for each component of the MI-EPEC lessons. These data included statements of purpose for evaluating each component and indicators of teacher behavior that represent the design criteria used to develop the lessons. The three

men

cons

items

elimi

scale

result

purpos

include

the MIT

three m

then use

additiona

containe

In

members

responses

The mean

MIT was

W

instructio

considera

EPEC K-

inservice e

to other u

members of the MIT provided 148 indicators. These indicators resulted in a pool consisting of 115 items.

Step 3 involved reviewing the appropriateness of the item pool and refining the items. The evaluation of the prototype items by the MIT resulted in items being accepted, eliminated, added and/or changed. Additionally, the format of the instrument and its rating scale were revised in accordance with suggestions provided by the MIT. This process resulted in a 64 item instrument.

In step 4, the MIT judged the appropriateness of the instructions in describing the purpose of the instrument, how it should be used and its format. The results obtained included additions, deletions and alterations in the instructions of the instrument.

In step 5, use and refinement of the instrument, variations in rating responses by the MIT were resolved by discussing the rationales underlying the discrepant scores. Over three meetings these discrepancies were resolved by item refinement. The instrument was then used a second time when criterion scores were established for each item and some additional changes were made in the instrument. At the end of step 5 the instrument contained 76 items.

In step 6, the content of the instrument was judged to be valid by the MIT members, and in step 7, reliability was established. For the instrument's continuous responses (ratings 1 to 5), the mean inter-rater reliability among all MIT members was .77. The mean correlation between the SRT members and the criterion scores established by the MIT was .52.

When this instrument is used by individuals knowledgeable of MI-EPEC instructional materials and who were involved in the development of the instrument, considerable promise for measuring the implementation behavior of teachers using MI-EPEC K-2 lessons was obtained. Continued refinement of the instrument, and the inservice education devoted to its appropriate use, may improve this promise and extend it to other user groups.

DEDICATION

To my great and wonderful father and, in particular, to my beloved, compassionate and wonderful mother who died just before I came to the United States to pursue my Ph.D., may God rest her soul in heaven.

To all of my brothers and sisters who taught me the sweetness of life.

To Dr. Paul Vogel who extended his support and understanding to me.

To all people and friends for their support and help.

ACKNOWLEDGMENTS

First, all thanks and praise to almighty God (Allah), the most merciful, the most compassionate, Lord of Lords and King of Kings, for his mercy and help in completing this work.

My work would not be accomplished without help and support of many people who contributed their time and effort.

I am very sincerely grateful to Dr. Paul Vogel, the chairperson of my doctoral guidance committee, for his advice, support, encouragement and understanding in my studies from the beginning when I started studying the English language, through my course work and ultimately in completing my research. Moreover I consider Dr. Vogel as wise, fair, knowledgeable and highly experienced; he knows very well how to treat others respectfully. Of course, thanks also to his predecessors: his parents, grandparents and all who came before.

Special thanks go to Drs. John Haubenstricker, Lynnette Overby and Steve Yelon, who served on my committee, for their professional support. Many thanks also go to Drs. Ray Allen and Rick Albrecht for their effort, time, support and assistance throughout study. Thanks also to Eileen Northrup, for her understanding, her help and support in editing the dissertation.

Thanks to Michigan Fitness Foundation for its support, facilities, equipment and materials wich contributed greatly to the conduct of the study.

I would also like to thank my parents, brothers and sisters for their help, support and encouragement in completing my Ph.D. degree.

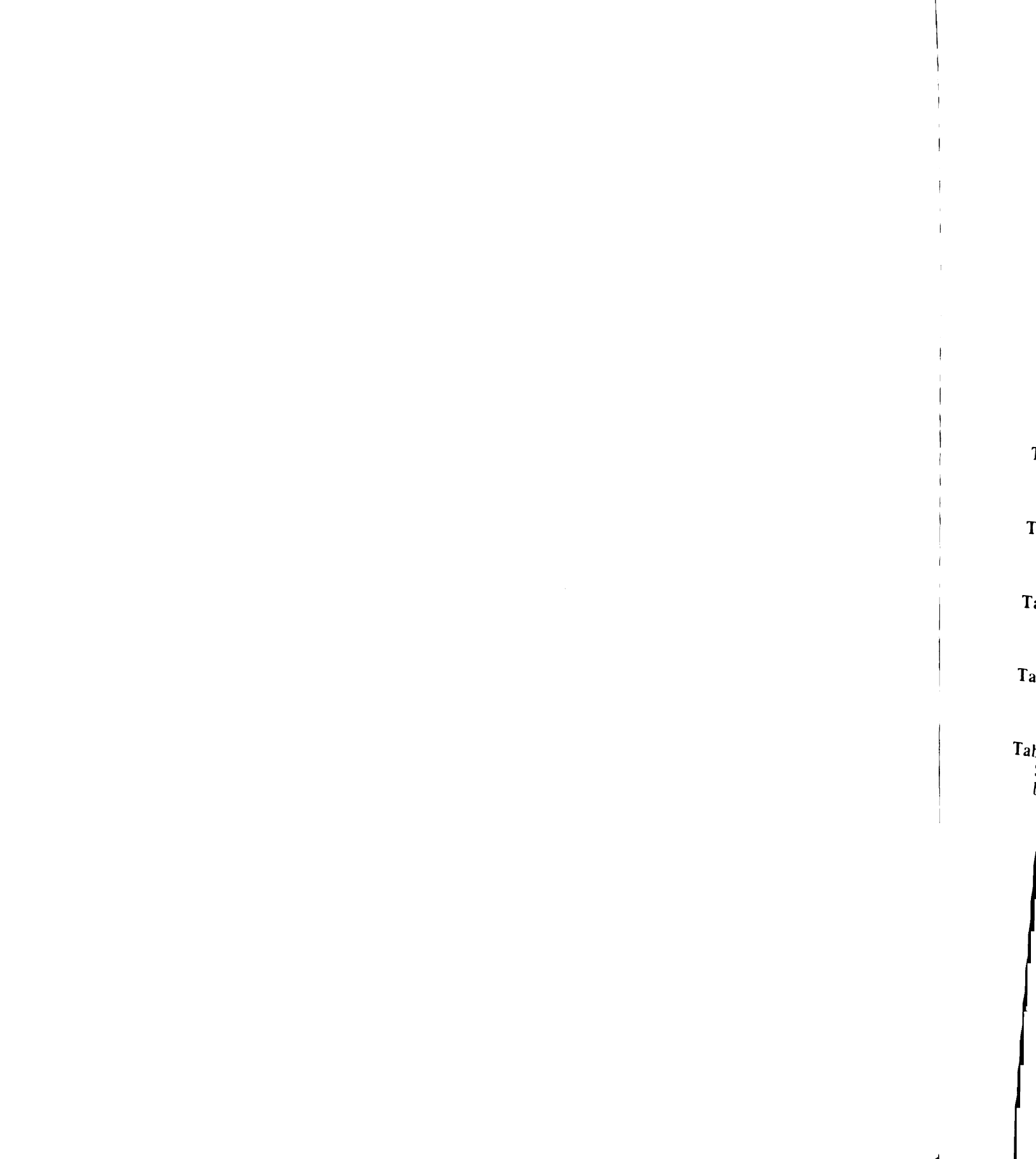
C

I.

II.

TABLE OF CONTENTS

LIST OF APPENDICES



Page

CHAPTER I

Introduction

Overview of the Problem

Physical activity has been associated with the prevention and control of several medical conditions that are major causes of death and disability in the United States. These conditions include coronary heart disease, hypertension, non-insulin-dependent diabetes mellitus, osteoporosis, obesity, musculoskeletal disease, high blood pressure and elevated blood fat (Lee et al., 1995; Harris et al., 1989; Morris et al., 1980; Lakka et al., 1994; Tenebaum, Singer, & Dishman, 1992; Nieman, 1989; Crews and Landers, 1987; and Seefeldt, 1986). Of these, coronary heart disease (CHD) is the leading cause of morbidity and mortality in the United States. It is estimated that an average of one in five persons will acquire CHD by the age of 60. Of primary importance is the fact that even modest amounts of physical activity have a protective effect in preventing this disease (Blair et al., 1989).

There are also positive effects of physical activity on psychological well being (Tenebaum, Singer, & Dishman, 1992; Nieman, 1989; Crews & Landers, 1987). For example, Tenebaum, Singer, and Dishman stated that there is a positive causal effect attributed to physical activity on self-esteem, mood, self-confidence and general awareness.

Current evidence makes it clear that appropriate use of physical activity can positively influence morbidity, mortality and the general quality of life. It is also clear that many children and adults are not benefiting from the effects of physical activity.

Estimates indicate that only one-fifth of the adult population is physically active (Center for Disease Control, 1987). Evidence also suggests that although there has been an increase in leisure-time physical activity by American adults during the last two decades, the rate of increase has recently declined.

Additionally, children in Michigan, when compared to their peers in other states, exhibit several elevated health risk factors that are precursors to serious diseases (physical

inactivity, obesity, elevated cholesterol and high blood pressure). Recent evidence also suggests that the youth of Michigan are not acquiring the motor skills necessary to enable physically active lifestyles (MEAP, 1984).

The benefits of physical activity on the quality of life and the need for Michigan residents to be more physically active are well supported in the professional literature. This support is confirmed by several prominent organizations in the form of strong position statements which advocate regular physical activity and the need for effective programs of physical education. These organizations include the:

- U.S. Department of Health and Human Services [Healthy people 2000: National health promotion and disease prevention objectives (1990)].
- American College of Sports Medicine [Opinion statement on physical fitness in children and youth (1988)].
- American Academy of Pediatrics [Opinion statement on physical fitness and the schools (1987)].
- American Heart Association [Position statement on benefits and recommendations for physical activity programs for all Americans (1992)].
- Centers for Disease Control and Prevention [Special communication on physical activity and public health (1995)].

Accordingly, increasing citizen levels of physical activity has become a major public health initiative.

The physical education curriculum in grades K-12 has the potential to interface with 97% of the children in Michigan. Although only modest amounts of instructional time currently is spent in physical education, this medium, more than any other initiative, provides the opportunity for the greatest number of citizens to acquire the beneficial effects of activity and the motor skills necessary to be fit for a lifetime. The physical education curriculum is uniquely responsible to enable young people to understand the importance of

physical activity and to obtain the fitness, knowledge, motor skills and personal/social attitudes necessary for wise use of activity for life.

Watson et al. (1994) said that if we can teach youth to be physically active, good things would happen. They advocate a curriculum that would help students gain skills and physical fitness as a result of regular physical activity. They also wanted to teach students to remain active in later years. Jordan (1993) stated that promoting active lifestyles lies at the heart of the health, physical education, recreation and dance professions. Whether we are teachers, practitioners or consultants we are all in the business of promoting active lifestyles for all individuals. Students also need to develop knowledge and values related to physical activity. It is a presumption, however, that participation in physical activities will produce graduates who possess the knowledge, skills, fitness and attitudes that will result in their being fit for a lifetime. The fact that 60% of the current population is sedentary suggests this is a false assumption. Surely more than 40% of the population participates in school physical education programs. Additionally, there is little evidence that current programs of physical education are producing their outcomes (Vogel, 1986).

It is clear that physical education has the potential to positively influence the health and well being of children for a lifetime and that children need such programs. It is also clear that physical education programs must be of high quality to provide the above stated benefits.

Status of Physical Education Programs

Physical education programs fail to meet criteria that define high quality. Lambert (1987) stated that, generally, poor scope and sequence, unbalanced curriculum (which focuses too much on sports), too many "exposure" activities in place of developing "competence", too much repetition of the same skills and the same sports are common problems within the physical education curriculum. Vogel and Seefeldt (1988) added that "many of the criticisms of physical education programs are related to their organization.

Among the common complaints are lack of sequential progression, redundancy of instruction, too much or too little content, activities unrelated to goals and objectives, content placed at inappropriate grade levels, and units that are too long or too short."

Vogel and Seefeldt (1988) pointed out that physical education programs are facing a continuing erosion in resources. These resources include time, staff, equipment, facilities and materials. An important reason for the erosion of resources is lack of administrative and public support. This lack of support is connected to the belief that participation in physical education programs does not result in identifiable student learning. This perception, even if it is not correct, has a restricting effect on needed resources. Ennis (1992) indicated that one of the most difficult problems that faces physical education teachers engaged in school-based curriculum design is the conflict between the large amount of information available about skills, sports and fitness, and the resource constraints that make it difficult to teach this content to students.

Designing high quality physical education programs requires use of specific criteria that clearly delineate the nature of their content, organization, materials and procedures. Vogel and Seefeldt (1988) define an exemplary physical education program as one that meets three broad criteria:

1. The intended outcomes of the program must be appropriate (defensible in terms of potential content and relevant in terms of stakeholders' values).
2. The program's organization and implementation procedures must be described sufficiently well so that they can be replicated.
3. There must be evidence of significant student achievement of the intended outcomes of the program.

Vogel (1998) added the following specificity to the above criteria in the introduction to the Michigan Exemplary Physical Education Curriculum Project (MI-EPEC) K-2 lesson materials:

1. Goals that quickly communicate the program's intended purpose and content that is scientifically defensible and socially relevant.
2. Program and instructional objectives that operationally define the meaning of the goals, are clearly stated, and are measurable.
3. A curriculum structure that identifies when, within grades (k-12), and for how much time, instruction should occur for all objectives.
4. A procedure that systematically provides for determining what objectives are included (and excluded) to meet resource constraints in local districts.
5. Instructional activities that are directly connected to the program and instructional objectives included in the approved curriculum.
6. Instruction that is consistent with the scientific literature on teaching and learning, and with the knowledge of professional practice.
7. Evidence that appropriate use of the program's materials and procedures result in student achievement of stated objectives.
8. A description of the curriculum and its implementation procedures that is sufficient to allow replication by others who desire to achieve similar effects.
9. Procedures for systematically refining program materials and procedures based on their in/effectiveness or in/efficiency that maximize changes based on strong rationales and minimize changes based on unsubstantiated personal bias.

Additional criteria for evaluating strengths and weaknesses of physical education programs are articulated in Dummer et al. (1993). The evaluation criteria included in this document (Dummer) were converted to design criteria to elaborate on the more general criteria stated above. They were then used to guide the development of Michigan's Exemplary Physical Education Curriculum (MI-EPEC).

The acceptance of the importance of physical activity, and knowledge of its potential benefits, have caused professionals in the field of physical education to initiate the design of exemplary physical education programs. One such effort is currently ongoing in

Michigan. The MI-EPEC project has been established to develop a physical education program that enables the youth of Michigan to obtain the necessary fitness levels, motor skills, knowledge and attitudes to be fit for life.

After developing design specifications, the structure of the program materials and procedures were developed to guide instruction on high priority program objectives. Subsequent to development of Instructional Resources Materials, lessons for grades K, 1 and 2 were written, field tested, revised and published in their first edition. At this point it is critical to evaluate the degree to which the lessons are implemented as intended and their effect on facilitating student achievement of stated objectives. To conduct evaluations of this type requires an instrument to assess implementation fidelity that is reliable and valid.

Purpose of the Study

By documenting implementation variability it is possible to establish the relationship between implementation and achievement of intended outcomes. Measurement of implementation fidelity is also necessary to evaluate the effectiveness of inservice/preservice training and to estimate the effects of various contexts, teacher characteristics and program content on program implementation and effectiveness. Because evaluation of implementation fidelity is critical to the evaluation of MI-EPEC and the development of quality programs, the intent of this study is to develop an instrument to measure implementation fidelity. The specific purpose of the study was to develop an instrument to measure the degree to which physical education teachers implement the K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project as intended by lesson developers.

Definitions

Michigan's Exemplary Physical Education Curriculum (MI-EPEC): A project of the Michigan Fitness Foundation. The mission of the MI-EPEC project is to enable those responsible for physical education in Michigan to create exemplary programs. This requires developing and disseminating materials and procedures that result in the youth of Michigan obtaining the necessary fitness levels, motor skills, knowledges and attitudes to be fit for life.

Implementation fidelity: The degree to which physical education teachers implement MI-EPEC K-2 lessons the way they were written by the lesson developers.

Delimitations

This Teacher Implementation Evaluation Instrument (TIEI) is designed to measure the implementation behavior of teacher using the MI-EPEC K-2 lessons devoted to the teaching of motor skills. Application of the instrument to cognitive, affective and/or fitness content were not included in the development of this version of the instrument. Additionally the instrument is designed to be used by evaluators trained in taking observational data from videotapes of teaching behavior to measure teacher implementation of lessons in grades K-2.

Limitations

Developing an instrument to measure the degree to which physical education teachers implement K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project as intended by lesson developers has the following limitation:

The instrument is not appropriate for assessing the fidelity of individuals teaching physical education with materials other than the MI-EPEC K-2 lessons.

CHAPTER II

Literature Review

Introduction

The intent of this chapter is to identify literature related to the development of an instrument for assessing teacher implementation fidelity of physical education lessons. Of particular importance is identifying theoretical guidelines for systematically developing an instrument suitable for measuring teacher implementation behavior.

To develop an instrument, it is necessary to examine the related literature to identify what others have done in this area and how it can be used in the current study. Sources used included documents available through Michigan State University's electronic library system (Magic), Dissertation Abstracts, ERIC documents and journals. Studies included in this chapter were selected for their potential contribution to the current study.

Many studies are available that address the development of instruments and many different variables and methods were discovered that fit the purposes of the various studies reviewed. However, there were also many similarities identified in the production of instruments.

This chapter is divided into four sections. Each section includes a review of topics thought to contribute to the development of the MI-EPEC instrument.

Section 1 overviews topics related to the systematic observation process as follows:

1. Nature of systematic observation
2. Use of systematic observation
3. Suitability of systematic observation
4. The process of systematic observation
5. Sources of observer error
6. Traditional methods of collecting observation data

7. Limitations of traditional data collection methods
8. Primary functions and limitations of systematic observation
9. What can be observed and measured in the classroom
10. Rating scales
11. Videotaping procedures
12. Training of observers
13. Test specifications

Section 2 covers the process (methodological steps) for developing a test and/or instrument.

Section 3 emphasizes steps used to establish validity.

Section 4 emphasizes establishing the reliability of an instrument.

Systematic Observation

Nature of systematic observation

Furst and Hill (1971) defined systematic observation as a set of procedures that use categories to code and quantify classroom behaviors of teachers and students. The procedures advocated require that observed behaviors be coded or classified according to relatively objective criteria that describe specific behaviors or actions.

Most systematic observation instruments are designed for research rather than for evaluating and measuring teacher performance in regular implementation settings. Simian and Boyer (1974) reported on 99 observation systems, of which all except one have been used in research. Twelve were reported to have been used for performance evaluation. Also, they reported that 52 of the 99 systems studied are being used as tools to give information directly to trainees. These observation systems act as mirrors for teachers and/or researchers to obtain feedback about classroom teaching behavior.

U

h
cı
in
m
in
(N

In
lea
to
the
fin
like
stu
fro
1.

2.

3.

4.

Use of systematic observation

Darst, Zakrajsek and Mancini, (1980) stated that systematic or direct observation has been used in areas such as anthropology, social psychology, clinical psychology and cross cultural psychology. Only since the 1960s has systematic observation become an important investigative procedure in the study of classroom teaching. Observational methodology is most appropriate when an argument can be made that some unique kind of information can be obtained by examining the behavior of a subject in a real life setting (Morra, undated).

Process-product relationships, connect teachers' behaviors and students' learning. In 1980 Brophy articulated a series of these relationships. For example, students tend to learn more when: their teachers believe that instructing students in the curriculum is basic to the teaching role; they expect their students to learn and act accordingly; they make it their business to see that students master key objectives of the curriculum by re-teaching or finding other ways to teach if the first approach is not successful; and they run business-like, task-oriented classrooms. Student learning was also connected to teachers keeping students engaged in meaningful tasks, and recognizing that students need direct instruction from the teacher. Brophy goes on to explain that:

1. Effective teachers minimize the time devoted to transitions and other purely procedural matters, and especially the time devoted to dealing with classroom disruptions.
2. Students are likely to remain attentive and engaged when their teacher presents an appropriate activity for them to focus on, keeps the activity moving at a good pace, and monitors their responsiveness.
3. Students seem to learn the most when they proceed rapidly but in very small steps especially when the teachers are available to give feedback.
4. Students taught with structured curricula do better than those taught with more individualized or discovery learning approaches.

5. Stud
thos
6. Stud
the t
7. In g
far a
struc
expl
resp
divic
frequ
supe
8. With
teach
9. The
field
but b
appre
10. Stude
recog
stude
11. Teac
"test
whe
peer
It
played a

5. Students who receive much of their instruction directly from the teacher do better than those expected to learn on their own or from one another.
6. Students in the early grades seem to require a lot of one-to-one dyadic interaction with the teacher, who gives them opportunities for practice and feedback.
7. In general, to the extent that students are younger, less intelligent, and/or are not as far along in mastering the key objectives of a given curriculum, teachers will need to structure their learning experiences, give more detailed and redundant instructions and explanations, interact more individually and more often with each student, elicit responses to questions and performance demands, provide individualized feedback, divide work assignments into smaller segments or devise ways to provide more frequent monitoring and corrective feedback, and, in general, continually direct and supervise learning activities.
8. Within any particular grade level, students who get more direct instruction from their teacher will learn more than students who get less direct instruction.
9. The teachers who are most successful with the students who are anxious, insecure or field dependent in cognitive style get top performance from them not by demanding it, but by fostering it gradually through praise, encouragement, expressions of appreciation for effort and shared pride and happiness for accomplishment.
10. Students will probably be better served in the long run if teachers are trained to recognize and respond appropriately to the needs and preferences of each individual student.
11. Teacher presentation of the same content tends to be more effective when they include "test-like events" that require each individual in the class to respond actively than when they merely require the majority of individuals to passively observe while a few peers respond.

It is safe to say that systematic observation, even with its limitations, has indeed played a major role in generating important guidelines for good teaching and has

contrib

played

area of

Suitabi

that mea

content t

behavior.

argument

the behav

observatic

that are in

observatio

arguments

instrument

have seriou

Dar

trained pers

interactions

would agree

observation

pointed out

react sensit

The proc

Th

1983).

contributed to the development of a language of teaching. Also, systematic observation has played a major role in the emergence of teaching/coaching behavior research as a bona fide area of empirical study in our profession.

Suitability of systematic observation

Systematic observation instruments developed initially for research include items that measure specific teaching skills, such as the method the teacher used to deliver the content to the students, distribution of instructional events, and other teaching learning behavior. Vogel (1988) said that observational methodology is most appropriate when an argument can be made that some unique kind of information can be obtained by examining the behavior of a subject in a real life setting. He added that, in general, any decision to use observational methodology must be accompanied by a careful consideration of the costs that are involved versus the need for this type of information, and he suggested that observational studies are most appropriate when: 1) small scale studies are used in validity arguments for developing less costly methods (i.e., questionnaires, self-rating instruments), and 2) studies are used which will provide data for making decisions that have serious consequences (i.e., hiring, promotion, graduation, etc.).

Darst, Zakrajsek and Mancini (1983) said that systematic observation allows a trained person following stated guidelines and procedures to observe, record and analyze interactions with the assurance that when others viewing the same sequence of events would agree with her/his recorded data. From this definition we can see that systematic observation includes both observing and recording. Johnston and Pennypacker (1980) pointed out that the goal of observation is to arrange conditions so that man or machine will react sensitively to the defined dimensions of the subject's behavior.

The process of systematic observation

The process of systematic observation involves the following steps (Siedentop, 1983).

1. Decidir

A d

observe

2. Develo

Hav

observe

behavic

definiti

the rou

on eithe

the topc

that ma

focus o

3. Selectir

Eac

the fact

often; o

events t

critical

develop

charact

4. Establ

Af

Reliab

collec

using

Obser

1. Deciding what to observe.

   A decision should be made in regard to the specific behaviors that need to be observed.

2. Developing definitions for the behaviors to be observed.

   Hawkins and Dobes (1977) stated that well-developed definitions make the observer's job of discriminating whether an event constitutes an instance of the target behavior much easier. Disagreement between observers can be minimized when definitions are clear, complete and objective. Barlow and Hersen (1984) said that once the rough outline is in place, the final definition is developed. This definition focuses on either of two aspects of the behavior; namely its topography or its function. When the topography of the behavior is emphasized, the definition describes the movements that make up the behavior. However, if function is emphasized, the definition needs to focus on the outcome or consequence of the behavior.

3. Selecting the most appropriate observation system

   Each behavior has two features: repeatability and duration. Repeatability refers to the fact that a behavior can occur over and over again. Some behaviors might occur often; others might occur only a few times. There is a duration to each behavior. Some events take a short time; others take a longer time. Both repeatability and duration are critical in the analysis of teaching and coaching environments. Thus, when you develop the definition of the behavior, you will decide whether the behavior is characterized by its frequency of occurrence or its typical duration.

4. Establishing observer reliability

   After selecting a system it is important to develop a sufficient level of reliability. Reliability is an important feature of the process because it is a prerequisite for collecting accurate data. Reliability is measured by the degree to which two persons using the same definitions and viewing the same activities agree on their coding. Observer reliability is often dependent on sound training of the observers.

5. M

to t

6. Summ

If

way a

purpos

When

impro

In

steps to t

5. Making the actual observations

There are some considerations the observer needs to be aware of when he/she goes to the actual setting where the observation is going to be made.

a. Subject reactivity is a definite possibility, especially if it is the first time that teachers or coaches are observed. If it is the first time for teachers to be observed, they might change their regular behavior because of the observer's presence.

b. There are a few things observers can do to minimize this reactivity. The following suggestions should be considered:

1) Teachers should explain to their students in general terms the purpose of the visit by the observer.

2) The observer should arrive early, so when the first students enter the observation may begin.

3) The observer should try to be as inconspicuous as possible, both in dress and behavior.

4) The observer should avoid interaction of any kind with the students.

5) If students are the target of the observer, the observer needs to ensure that he or she disguises the attention by varying his or her glances.

6) Try to keep equipment out of sight as much as possible.

6. Summarizing and interpreting the data

If data are collected for supervisory purposes, they need to be summarized in such a way as to provide feedback to the instructor. If they are collected for research purposes, they need to be converted into values appropriate for statistical analysis. When the data are collected for program evaluation purposes, they should be used to improve the program through the refinement process.

In addition to the previous steps, King, Lyons and Fitz-Gibbon (1987) added these steps to the process of formal observation:

1. Prepar

2. Decid
determ

3. Prepar

4. Prepar

5. Choos

Sources

Th
records.

1. Behavi
behavi
develo
constit
observ

2. Resear
For ex
with sh
the obs

3. Proble
observ
record
reliabl
If
support th
A
five majo

1. Prepare scenarios of episodes that should not occur.
2. Decide how long each observation time sample must be in order to yield good data and determine how many time samples are needed.
3. Prepare a sampling plan for conducting observations.
4. Prepare the observers' recording sheets.
5. Choose observers.

Sources of observer error

There are three factors that influence the accuracy and completeness of the coding records.

1. Behavior definition is the degree to which the observers can understand or interpret the behavior being observed in the same way. Hawkins and Dobes (1977) stated that well-developed definitions make the observer's job of discriminating whether an event constitutes an instance of the target behavior much easier. Disagreement between observers can be minimized when definitions are clear, complete and objective.
2. Research design is the type of research plan used to observe and record the behavior. For example, if the research requires the observer to conduct pre- and post-observation, with short periods of time between (less than one week), it will affect the accuracy of the observations.
3. Problems of measurement are related to more than one concept. For example, if the observers do not position themselves properly, they will not be able to observe and record the behavior accurately. Also, if the observation instrument is not valid and/or reliable, the observation and the data recorded will not be accurate.

If these problems are ignored, the data will not provide sufficient evidence to support the record of teacher behavior.

Another factor that can influence accuracy is the nature of the observers. There are five major obstacles that relate directly to the observer.

1. Observ

Ac

change

called

as goin

with th

neatly

boredo

a.

b.

c.

2. Compl

Th

catego

of char

comple

the len

3. Observ

If

might

behav

even

minin

a

b

c

1. Observer drift

According to Johnson and Bolstad (1973) observers' drift refers to the tendency to change coding rules and interpret category definitions differently across time. It is also called instrument decay. Observer drift is a gradual process that has many causes, such as going for a long period without using a system, mixing definitions from one system with those of another, and observing individual instances of behavior that do not fit neatly into any one category and making accommodations for that. Satiation and boredom also cause drift tendencies. Procedures to identify the observer drift include:

a. Continuation of observer training.

b. Periodically coding prescored videotapes and checking scores with the criterion scores.

c. Periodically rotating the observer with whom coding is compared.

2. Complexity of the observation system

The difficulty in making correct coding decisions increases as the number of categories increases. The observers may have a hard time keeping up with the fast pace of change in behavior patterns. Based on the observation system used, one can reduce complexity by reducing the number of categories that are coded or slightly increasing the length of the intervals.

3. Observer expectancies/bias

If observers are told that changes in the behavior of teachers, coaches and students, might occur as a result of some type of intervention, coupled with feedback about their behavior pattern, they are more likely to produce bias in the data that reflect a change even if the behavior did not change. Kazdin (1977) suggested that observer bias can be minimized by:

a. Videotaping the sessions.

b. Frequently inserting new observers.

c. Keeping observers naive as to the purpose of the investigation.

4. Obse

B

the pu

contr

tend t

5. Obse

B

alterat

raw d

a.

b.

c.

d.

Traditio

Sy

collecting

Traditiona

Each of th

in preserv

Eyeballin

and viewi

he/she sa

observed

things th

4. Observer reactivity

Bolstad (1973) said that knowing that someone else is also doing an observation for the purpose of checking reliability is called observer reactivity, and that it is hard to control this factor. When observers know they are being tested, agreement percentages tend to be higher.

5. Observer cheating

Barlow and Hersen (1984) stated that cheating occurs by way of data fabrication, alteration of data, and incorrect calculation of agreement scores or derivatives from the raw data. Cheating opportunities can be minimized by:

a. Collecting the coding sheets immediately after an observation session is completed.
b. Using pens rather than pencils.
c. Letting persons other that the observers do the calculation of agreement percentages.
d. Conducting unannounced, random reliability checks.

Traditional methods of collecting observational data

Systematic observation continues to compete with other more traditional methods of collecting information on what teachers, coaches and their students or athletes do. Traditional methods include eyeballing, anecdotal recording, rating scales and checklists. Each of these techniques has played a major part of the evaluation process of teachers, both in preservice teacher preparation programs and in school districts.

Eyeballing refers to anyone from the outside, like an administrator, entering a classroom and viewing the ongoing activities without making any formal written record of what he/she saw. Based on the memory of what was seen, he/she may provide feedback to the observed teacher. Anecdotal recording involves the observer writing down some of the things that he/she sees and hears.

Limitations

Some

1. Lack of th

The s

Johnston

poor obs

experien

importan

what act

2. Reliabili

Ther

another

4 is equ

widens

indepen

often th

and ex

lacks r

3. Specif

T

on w

are s

The

ofte

teac

fol

Limitations of traditional data collection methods

Some of the limitations of traditional data collection methods are:

1. Lack of the objectivity

The source of data comes from observer's opinion about what events occurred. Johnston and Pennypacker (1980), and Siedentop (1983) noted people are notoriously poor observers. Opinion is typically based on personal biases and history of experiences. Thus, the observer has a strong tendency to report what he/she thinks is important. Consequently, the resulting record is what he/she wants to see, rather than what actually happened. Thus, rating scales often lack objectivity.

2. Reliability

There is no way to determine why a person viewing a behavior scores it a 3 and another person viewing the same event scores it a 4, or if the difference between 3 and 4 is equal to the difference between 4 and 5. This becomes a problem when the scale widens from a 2-point scale to a 5- or 7-point scale. Also, if we ask two or three independent observers to evaluate a teacher using anecdotal recording, the result is often three different narrative records. This variability can be attributed to personal bias and experience. Accordingly, data collected through anecdotal recording most often lacks reliability.

3. Specificity

The information provided through rating scales cannot provide a specific database on which recommendations for improvement can be made. Rating scales and checklists are simply too crude to show that improvements were actually made and to what extent. The point of identifying these limitations is to stress that the traditional methods are often unreliable ways of collecting data within the context of analyzing and evaluating teaching performance, primarily because there are no strict rules and procedures to follow.

Primary fur

Syste

research and

information c

supervision c

1. Adminis

2. Staff dev

3. Coopera

during th

4. The univ

Syst

1. It conce

2. Users n

system

objecti

3. Descri

could

evide

teach

4. Findi

word

cons

What

instruc

behavi

Primary functions and limitations of systematic observation

Systematic observation tools play a major role in two different but related areas: research and supervision. As part of research projects, systematic observation provides information on both independent and dependent variables. Locke (1979) stated that supervision can serve a variety of purposes. It plays a role in:

1. Administrative decisions in public schools regarding the retention of teachers.
2. Staff development programs in public schools.
3. Cooperative efforts among teachers to guide new, inexperienced student teachers during their internship.
4. The university supervisor's contributions in that same setting.

Systematic observation has several crucial limitations:

1. It concentrates only on observable events and behaviors.
2. Users need to be aware of the fact that, when used appropriately and reliably, systematic observation produces only descriptive information that is relatively objective.
3. Descriptive data in and of themselves cannot give prescriptions as to what a practitioner could or should change. Hawkins, Wiegand and Landin (1985) said that there is evidence that data-based feedback to teachers frequently consists largely of reporting to teachers what happened (much like knowledge of results).
4. Findings obtained through systematic observation are always contextual. In other words, the message they may provide about teaching performance needs to be considered in light of the situation in which they were observed.

What can be observed and measured in the classroom

Rink (1979) said that the purpose of observation instruments in physical education instruction is to describe the process of content development. She categorized teaching behaviors into three major categories:

1. Commun

This cate

categoriz

a. Solic

Any

anoth

educ

that

b. Resp

Any

situa

answ

c. Initi

Any

dire

beh

wit

son

d. Ap

An

as

bo

2. Conte

The f

deve

1. Communication function

   This category focuses on the type of communication used for a given behavior, and it is categorized into four sub-categories:

   a. Soliciting

      Any spoken or unspoken behavior that causes, or intends to cause, a behavior from another person in the instructional situation. Soliciting behaviors in physical education generally involves giving commands to do something or asking questions that require a response.

   b. Responding:

      Any spoken or unspoken behavior that is a response to an event in the instructional situation. Responding behaviors in physical education setting are generally answers to verbal questions.

   c. Initiating:

      Any spoken or unspoken behavior that represents information to others that is not directly elicited by an antecedent behavior or event. An example of initiating behaviors in the physical education settings is lecture (providing information without expecting an immediate response). Directions on how or why to do something would also typically fall into this category.

   d. Appraisal:

      Any spoken or unspoken behavior that directly communicates a judgment or assessment about performance and is not a corrective statement that solicits future behavior.

2. Content function

   The function of this category is to describe the role the behavior plays in the development of lesson content, and it is categorized into these six sub-categories:

a. Inf
Ar
inf
fu
co
b. Re
Ar
su
fe
ho
pe
c. Ex
Ar
ex
or
d. Ap
Ar
act
pe
in
e. C
A
o
b
f. C
A
a

a. Informing:

Any spoken or unspoken behavior that is intended to communicate substantive information to the learner and does not have a refining, extending, or applying function. Informing behaviors do not develop content; they merely introduce content information to the learner.

b. Refining:

Any spoken or unspoken behavior that is qualitatively related to improving substantive motor performance. Refining behaviors can be solicitations (corrective feedback), appraisals (specific evaluation of performance), initiations (lecture on how to improve performance), or responses (answers to a question on how to perform better).

c. Extending:

Any spoken or unspoken behavior that is quantitatively related to either reducing or expanding the content of material. Extending behaviors usually make things more or less difficult, or are related to a variety of responses.

d. Applying:

Any spoken or unspoken behavior that introduces a focus on the use of motor activity that is related to the use of that movement rather than the movement performance itself. Application content in physical education is usually competitive in either a self testing or game situation.

e. Conduct:

Any spoken or unspoken behavior that structures, directs or reinforces the explicit or implicit behavior code of a given situation. Conduct behaviors are disciplinary behaviors.

f. Organization:

Any spoken or unspoken behavior that structures, directs, or reinforces the arrangement of people, time or equipment to create conditions for substantive

learni

catego

3. Source/re

The sourc

student is

to reflect

category

group).

over thre

individu

Rating sc

Rati

using ratin

learning en

used or we

Go

procedure

characteri

other char

from syst

the latter

R

measure

defined,

reliabilit

scales is

judging

learning. Organizational behaviors are management behaviors without the conduct category.

3. Source/recipient of behavior function

The source of behavior is coded differently depending on whether the teacher or a student is the source of the behavior. If the source is the teacher, the category is coded to reflect to whom the teacher is directing the behavior. If the source is the student, the category is coded to reflect the nature of that source (a single student, class, or small group). A single student is one or two students, a class is whole class, a small group is over three students but not more than half the class, and individual/public is an individual but in a manner that the whole class can hear.

Rating scales

Rating scales and checklists are popular for describing teachers' performance. In using rating scales, observers record their opinion on various aspects of the teaching-learning environment. When checklists are used they mark whether certain things were used or were attended to by the teacher. Again scores are based on opinion.

Good (1959) defines ratings as an estimate, according to some systematized procedure, of the degree to which an individual person or thing possesses some characteristic. He defines rating scales as devices used in evaluating products, attitudes or other characteristics of instructors or learners. Rating scales used for observation differ from systematic observation instruments in that the former record general impressions and the latter describe in detail what has happened.

Rating scales are used more extensively than any other device in attempts to measure teaching competency. Often rating scales include items that are not operationally defined, do not allow for proper time sampling, and do not have evidence of inter-rater reliability. McNeil and Popham (1973) point out that a major deterrent to the use of rating scales is the failure to control for sampling. Remmers (1963) suggests five criteria for judging the appropriateness of rating scales as measuring devices.

1. Obj
unbi
2. Reli
erro
3. Sens
4. Valic
the d
5. Utilit
rate.

Videota

W
a curiosity
contexts i
increased,
research s
"knowledg
social inte
believe tha
presents li

Br
objective
important
or parapr
particular
objective
academic
intervent

1. Objectivity. Use of the instrument should yield verifiable, producible data that are unbiased by the peculiar characteristics of the rater.
2. Reliability. The instrument should yield the same values, within limits of allowable error, under the same set of conditions.
3. Sensitivity. It should yield fine distinctions about the object of investigation.
4. Validity. The instrument's content, in this case the rating scale items, should relate to the defined area of investigation and relevant constructs.
5. Utility. The instrument should be manageable to allow data collection at a reasonable rate.

Videotaping procedures

Wiemann (1981) said that, in the past decade, videotape technology has gone from a curiosity in the behavioral sciences to a stable research tool. Although the research contexts in which it is used and the problems upon which it is brought to bear have steadily increased, little attention has been directed to the potential reactivity of videotaping in the research situation, whether in the laboratory or the field. The potential reactivity of the "knowledge of being videotaped" is particularly crucial in studies that deal with "everyday" social interaction, as exemplified in conversation analysis. Most researchers seem to believe that observing and recording behaviors normally (not under conscious control) presents little danger of observer interference with the phenomena of interest.

Broome and White (1995) stated in their article that videotape feedback provides an objective recording of students' behaviors and classroom events, and therefore, is an important resource for assessment and instruction. To get started, they suggest the teacher, or paraprofessional, set the video camera on a tripod and record student responses to a particular instructional lesson in a particular classroom setting for a planned instructional objective. For example, the camera is purposely focused on a small group receiving academic instruction, a group counseling session, or the time-out booth during crisis intervention.

Anot

periods of th

Occasionally

for instructi

recommende

for teachers

students wit

be used to l

and other re

orientation

permission

Ph

select vide

appears to

procedure

stability a

of the vid

O

finding th

to watch

clearly. T

the video

recomme

operation

Such hin

taping st

close to

Another taping option is to simply run the tape at a wide angle during various periods of the school day without any particular instructional objective in mind. Occasionally these tapes catch helpful glimpses of student behavior that can be used later for instructional purposes. Gunter, Jack, Shores, Carrel and Flowers (1993) recommended review of videotapes of instructional interactions with students as a medium for teachers to evaluate their use of critical instructional and management behaviors with students with emotional and behavioral disorders. Also, the first videotape session should be used to let the students become accustomed to the camera. This also defuses "clowning" and other reactive behavior that students are apt to show initially. After the initial setup and orientation are completed, no further attention is needed. Before videotaping, parental permission should be obtained and it should be in writing.

Philip and Thomas (1996) stated that information for teachers regarding how to select videotape equipment and how to record high quality videotapes in classroom settings appears to be extremely limited. They do, however, include some tips for videotaping procedures. For example, they state that the camera should be placed on a tripod for stability and safety, and they provide a sample letter for use in informing parents/guardians of the videotaping procedure.

Often the first two or three recordings will be experimental, for the purpose of finding the best camera settings and classroom setup for recording videos that are pleasing to watch and that provide visual and auditory information that can be seen and heard clearly. The effort that goes into such experimentation gives students time to desensitize to the videotaping process. Also, they suggested that teachers rely heavily on the recommendations in the operational manual for their video recorder. For example, the operation manual for the "Sharp videocam" has a section titled "Hints for Better Shooting." Such hints as having the light source behind the recorder means that you should avoid taping students who are sitting in front of a window. You also need to place the camera as close to the selected group as possible; this will help to screen out background noise.

The p

that allows fi

this should i

behaviors fo

implementat

checked and

lessons.

Observer

Bar

in the use o

phases of l

the investig

less time c

successive

techniques

Phase 1: C

New c

includ

Phase 2:

Obse

discr

many

be c

shou

Phase 3

It is

suc

The primary requirement is that videotaped sessions provide a record of behavior that allows functional assessment. If teachers simply record entire instructional periods, this should insure that enough interaction has been captured to allow an adequate number of behaviors for calculations. To make sure that the process of videotaping the implementation of a lesson is done properly, specific procedures should be developed to be checked and/or followed by the physical education teachers who are going to implement the lessons.

Observer training

Barlow and Hersen (1984) stated that proper training facilitates observer reliability in the use of an observation system. This section provides an overview of the major phases of learning how to use a systematic observation system. Based on the purpose of the investigation and the complexity of the observation system, some of the steps may be less time consuming than others. They suggest using instructional principles such as successive approximations in teaching observer trainees appropriate observation techniques. For example

Phase 1: Orientation to the system

New observers should be introduced to the basic purpose of the observation system, including descriptions of the types of events or behavior to be studied.

Phase 2: Learning the categories

Observers should learn all the definitions of the categories. They should be able to discriminate among the basic categories with 100% accuracy. It is good to show as many videotaped examples of them as possible, making sure that each category cannot be confused with other closely related categories. Successful completion of this phase should be based on passing a written or oral test.

Phase 3: Using the coding form correctly

It is critical for observers to know how to use the coding form. It takes practice to successfully place the coding symbols in the appropriate areas of complex forms.

Phase 4: In

Observe

while vi

appropr

Phase 5: Li

Live ob

Observa

patterns

Test spec

In h

proportiona

(1976) state

1. Require

2. Allow

3. Use ite

4. Have t

5. Requir

6. Be usa

variou

Th

devices.

1) Object

categorie

Utility [t

Phase 4: Initial coding practice

Observers need opportunities to practice coding by using part of the observation tool while viewing a videotape of a physical education lesson or practice session. Using the appropriate videotape is critical at this stage; no single tape fits each observation tool.

Phase 5: Live observation practice

Live observational practice is needed in the place where motor skill instruction occurs. Observation sites should be chosen with care in terms of the complexity of behavior patterns to be coded.

Test specifications

In his study aimed to develop a paper-pencil test of the Piagetian levels of proportional thinking of junior high school pupils in the context of physical science, Ruud (1976) stated that the test should have the following specifications:

1. Require a 30-minute testing session.
2. Allow for the measurement of large numbers of persons.
3. Use items with different science content.
4. Have the reliability offered by several measures of the same person.
5. Require no expertise of the test administrator.
6. Be usable as a source of information for determining the numbers of pupils at the various proportional reasoning levels and which pupils are at each of these levels.

There are five criteria for judging the appropriateness of rating scales as measuring devices. These criteria are stated by Remmers (1963) as follows:

1) Objectivity, 2) Reliability, 3) Sensitivity, 4) Content [in this case the rating scale categories should relate to the defined area of investigation and relevant constructs], and 5) Utility [the instrument should be manageable to allow data collection at a reasonable rate].

Steps in t

Steps for d

More

instrument c

that should

1. Construc

test con

the test

develop

assigne

2. Decide

test que

combi

know

3. Const

4. Deter

T

1. Kee

2. Avo

3. Kee

4. Hav

5. Do

6. Ha

7. Av

8. Do

9. Ha

Steps in the Development of an Instrument

Steps for developing a general instrument

More than one investigator discussed the general steps necessary to develop an instrument or a test. Construction of a knowledge test requires several basic procedures that should be followed. General steps for the construction of knowledge tests are:

1. Construct a table of specifications. A table of test specifications provides an outline for test construction and ensures that all of the material covered in the course is included in the test and that the correct weight is given to each topic. The table should be developed after examining the objectives of the course and the relative importance assigned to each objective.
2. Decide the nature of the test. Generally, a written test should consist of several types of test questions. The types of questions will be based on teacher preference but a combination of questions are thought to give more complete evaluation of student knowledge.
3. Construct the test items.
4. Determine the test format and administrative details.

The construction procedures for test items include:

1. Keep the statement short.
2. Avoid trivial items.
3. Keep the vocabulary simple.
4. Have only a single idea in each statement.
5. Do not take statements directly from textbooks.
6. Have answers randomly placed.
7. Avoid the use of negative statements.
8. Do not have over 60% of the answers the same.
9. Have false statements that are plausible.

10. Avoid st

"all," "n

Steps

by Horine (1

1. Determir

2. Select te

3. Review

4. Prepare

items.

5. Conduc

prepare

6. Determ

7. Develo

Se

goals, the

necessary

developm

1. Esta

2. Cre

3. Sel

4. Iter

5. Co

6. Ite

7. Ite

8. Co

9. Ite

10. M

10. Avoid stereotyped determiners such as "always," "never," "sometimes," "usually," "all," "none" or "often."

Steps for developing a test more specific to a current teaching situation are provided by Horine (1981) and DiNucci (1988). They include:

1. Determine the purpose and define the test objectives.
2. Select test items to measure the specific traits to be measured.
3. Review the test literature to help in item selection.
4. Prepare test instructions, including specific test area layout instructions and order of test items.
5. Conduct pilot testing using a group of the same age and sex for which the test is prepared.
6. Determine validity, reliability and objectivity.
7. Develop standards for the specific population.

Seefeldt (1990) agreed with the previous two investigators in that establishing the goals, the specific objectives, and the weight or relative importance for each objective are necessary steps for the development process of a test. He added the following steps for the development of the Michigan Educational Assessment Program (MEAP) test:

1. Establish a coordinating committee.
2. Create a test blueprint, item prototypes, item specifications.
3. Select and train item writers.
4. Item writing.
5. Conduct conceptual reviews.
6. Item editing.
7. Item tryouts.
8. Content reviews.
9. Item bias reviews.
10. Measurement reviews.

11. Iter

12. Tes

13. Dis

F

junior hig

followed

testing ph

pencil tes

final item

Steps fo

O

justificati

need to u

for the ef

1. Have

2. Have

your e

3. Have

4. Have

5. Have

a. R

b. V

6. Have

N

1. W

he

11. Item pilot.
12. Test pilot.
13. Dissemination.

For developing a paper-pencil test of Piagetian levels of proportional thinking of junior high school pupils in the context of physical science, three major steps were followed (Ruud, 1976). They included: 1) initial trial or pilot phase, 2) an intensive task testing phase with 40 pupils to produce an initial item design, and 3) an extensive paper-pencil testing phase with groups that, in some cases, exceeded 300 pupils from which the final item set was written.

Steps for developing an observational instrument for assessing a program

Observational studies are generally so expensive that they require considerable justification to fit within the constraints of most evaluation budgets. But, for those who need to use this method, there is a list of steps, written as questions, that will help to plan for the effort (Morra, undated).

1. Have you carefully specified the uses to which this data will be put?
2. Have you specified the possible kinds of information that would yield clear answers to your evaluation questions?
3. Have you specified the format for recording observational data?
4. Have you made plans to train the observers?
5. Have you planned for the technical arguments to support these data?
   a. Reliability.
   b. Validity.
6. Have you planned for the analysis of the data?

Morra suggests some additional aids to help complete this plan.

1. Who will be responsible for the collection of this data, and how much time should he/she spend on it?

2. W
m
3. W
4. H
5. H
6. If
tr
7. W
8. H
9. W
ob
10. W
re
11. Ho
co
res
12. Wh
res
13. Wh
eva
In
of the co
using th
1. Wha
2. How
3. Wha
4. How

2. What staff members should work on the project, what will their duties be, and how much time they will spend?
3. When will the observers be selected and trained? (Who will be responsible?)
4. How will the observation sites be selected? (Who will be responsible?)
5. How will the observations be scheduled?
6. If you decided to use film or videotape: Is the equipment available? Do you have a trained person to operate it? (Who will be responsible?)
7. When, where and how will printed materials be produced?
8. How will the quality of this effort be maintained?
9. Who will be responsible for physically collecting the response forms from the observers?
10. Will the observers be compensated? How will this be done? Who will be responsible?
11. How and when will the analysis be conducted? If special resources such as consultants are required, have you made provisions for these? Who will be responsible?
12. When and to whom will the technical support arguments be made? Who will be responsible?
13. When and how will the final report be prepared, and how will it be used in making evaluation decisions? Are sufficient clerical resources available? Who is responsible?

In the development process of an instrument to measure the observable performance of the competence of preservice teacher education students, Shearron (1976) suggests using the following questions:

1. What is to be measured?
2. How specifically are the competencies stated?
3. What context defines the demonstration of competence to be measured?
4. How often and at what points should preservice teachers be measured?

5. Who will
measuren
6. Will the r
7. How muc

The I

(IOTAH), w

1. Introduc
complet
the time
2. Training
and con
develop
develop
observa
the inst
3. Operat
segme
suffici
obtain
for uti
organ
an op

Ic

of obser

A

But, ther

determin

5. Who will participate in measuring performance; who is going to be involved in the measurement process?
6. Will the measurement be formative, summative, or both?
7. How much time will be required, in adapting or developing measurement instrument?

The Instrument for the Observation of Teaching Activities in Higher Education (IOTAH), was divided into three stages (Carpenter, 1977).

1. Introducing the program (orientation): Based on the desired inclusiveness, completeness of the orientation desired, and the number of participants in the activities, the time (days and hours) for introduction activity will be specified.
2. Training (workshop): This stage includes an intensive training program by the director and consultants, activities that require the participant to become familiar with developing an understanding of an acceptable definition of teaching competence, developing objective observation skills through film training and actual classroom observation; developing skills in data recording, and making an accurate application of the instrument for the observation of teaching activities in higher education.
3. Operational phase (implementation): The implementation phase is the most vital segment of the program. It is undertaken after an organization has benefited from a sufficient number of workshops. Up to this point professional growth has been obtained mostly on the part of individuals who have attended workshops. Procedures for utilization of the program are developed cooperatively, additional training of the organization's staff is given, guidelines are established for implementing IOTAH, and an operation program benefiting all in the organization is formulated.

IOTAH has proven its reliability with trained observers. The inter-rater correlation of observations is around 90%.

As stated earlier, there are different steps involved in developing an instrument. But, there are more similarities than differences in the development process. For example, determination of the purpose and objectives of an instrument or test is common to all test

development

Jochuas et al

development

School of B

instrument s

used can be

1. Devel

2. Revie

3. Use a

4. Defin

5. Cond

6. Cond

7. Cons

8. Deve

9. Scre

10. Asse

11. Ana

12. Dev

13. Def

14. Fiel

15. An

16. Pu

T

develop

1. Det

2. Co

dev

development procedures and Fletcher and Spady (1975), Lopatka (1978), McDavis (1976), Jochuas et al. (1979) and Seidman et al. (1979) used the same and/or similar steps in the development of their instruments. Also, the methodology developed and applied in the School of Business Administration, San Diego State University, used steps to develop their instrument similar to those steps used by the investigators listed above. The steps that were used can be summarized as follows:

1. Develop a statement of need and purpose for the instrument.
2. Review literature to identify objectives the instrument is to fulfill.
3. Use a panel of experts to judge the appropriateness of each objective.
4. Define the areas of concern.
5. Conduct a content analysis of instruments that may meet needed concerns.
6. Conduct a process analysis of each instrument.
7. Construct and revise individual checklists.
8. Develop an item pool.
9. Screen the item pool.
10. Assess the screened items for importance and for user ability to rate.
11. Analyze response.
12. Develop evaluation instrument options.
13. Define evaluation models.
14. Field test items.
15. Analyze field testing items.
16. Publish the instrument.

To accomplish the purpose of this study, the following steps will be used in the development process:

1. Determine the purpose and objectives of the instrument.
2. Construct a table of test specifications to guide the instrument development process and develop a pool of items.

3. Review

4. Prepare

5. Use and

6. Establis

7. Establis

Validity

Val

what it is i

to which th

to describe

investigato

who stated

1. Conter

Th

validit

of a m

wheth

conce

the re

whetl

meas

adeq

unde

popu

indi

thes

3. Review and revise the item pool.
4. Prepare instructions for using the instrument.
5. Use and refine the instrument.
6. Establish content validity of the instrument.
7. Establish reliability of the instrument.

Validity

Validity of a test is defined as a measure of the degree to which the test measures what it is intended to measure (Ruud, 1976). Validity has also been defined as the degree to which the measures obtained by an instrument actually describe what they are supposed to describe (Wittrock, 1986). There is more than one type of validity. Different investigators discussed different types of validity, such as Nachmias and Nachmias (1976), who stated the following three types.

1. Content Validity

   There are two common varieties of content validity: face validity and sampling validity. Face validity rests on the researchers' subjective evaluation as to the validity of a measuring instrument. In practice, face validity does not relate to the question of whether an instrument measures that which the researcher wishes to measure; rather it concerns the extent to which it measures that which it appears to measure according to the researcher's subjective assessment. The primary concern of sampling validity is whether a given population of situations or behaviors is adequately sampled by the measuring instrument in question. That is, does the content of the instrument adequately represent the content population of the property being measured? The underlying assumption of sampling validity is that every variable has a content population consisting of an infinite number of items (statements, questions or indicators) and that a highly valid instrument constitutes a representative sample of these items.

2. Empi

T

instru

valid.

the in

relatio

Of the

most v

Pre

as a cr

words,

with ar

obtaini

obtaini

college

measur

3. Constru

Cor

framew

theoret

In a

and predic

Kazdin (1

measures

events. In

extent to

validity a

2. Empirical Validity

The concern of empirical validity is with the relationship between the measuring instrument and the measurement results. It is assumed that if a certain instrument is valid, then there should exist certain empirical relations between the results produced by the instrument and other properties or variables. Evidence to support the existence of a relation is obtained by measures of correlation appropriate to the level of measurement. Of the various tests designed to evaluate empirical validity, predictive validity is the most widely used.

Predictive validity is characterized by prediction to an external measure, referred to as a criterion, and by checking a measuring instrument against some outcome. In other words, predictive validity is the correlation between the results of a given measurement with an external criterion. For example, one can validate an intelligence test by first obtaining a set of test scores on a group, such as college freshmen, and then by obtaining the grade-point averages that these freshmen made during their first year of college. A correlation coefficient is then computed between the two sets of measurements. The obtained correlation is usually called the validity coefficient.

3. Construct Validity

Construct validity involves relating a measuring instrument to an overall theoretical framework in order to determine whether the instrument is tied to the concepts and theoretical assumptions that are employed.

In addition to the three previous types of validity stated above, (concurrent validity and predictive validity) can also be used for measuring the validity of an instrument. Kazdin (1977) added that it is critical that users of systematic observation tools take measures to ensure that the observation instrument they select provides a valid reflection of events. In the context of systematic observation, validity (also called accuracy) refers to the extent to which an instrument measures what it is supposed to measure. He defined validity as: The extent to which observations scored by an observer match those of a

predetermi

goal for th

as they occ

An

demonstrat

instrument

validity m

validity, c

different i

Ru

used conte

time, these

instrument

of Piagetia

physical s

their resol

when the t

initial test

expected

validity if

does not

D

and othe

instrume

importa

studies t

predetermined standard for the same data. Johnston and Pennypacker (1980) said that the goal for the measurements is to approximate as closely as possible the true value of events as they occur in the environment.

Any assessment instrument must be valid for its purpose. If the instrument demonstrated is not a valid instrument, the results will be questionable. The validity of the instrument can be judged through more than one method. Studies stated that specific validity methods, like content validity, concurrent validity, construct validity, discriminate validity, convergent validity, and face validity, have been used to judge the validity of different instruments.

Ruud (1976), Voelker and Horvat (1974), Colbert (1977) and Holland (1986) have used content validity as a method to judge the validity of their instruments. At the same time, these researchers also have used additional descriptions of the validity of their instruments. Content validity was used in a study conducted to develop a paper-pencil test of Piagetian levels of proportional thinking of junior high school pupils in the context of physical science. A test has content validity if the items in the test require behaviors for their resolution that are appropriate to the trait being measured. Concurrent validity exists when the test correlates highly positively with direct test measures of the same traits as the initial test and construct validity suggests that the difficulty level of items would be expected to show increasing difficulty with higher levels of the test. A test has construct validity if it measures the attribute it is intended to measure. It follows then that if the test does not measure other things, it is acceptable.

Discriminate validity is evident if a test discriminates between the trait it measures and other traits. Convergent validity exists when measurements correspond to other instruments measuring the same trait (Ruud, 1976). Content validity is one of the most important types of validity used to judge tests and/or instruments. Examples of some of the studies that used content validity follow.

H

fundamen

learning d

was used

for determ

judged co

The panel

gave their

C

(1976). I

Observati

process,"

LoUOI a

the Level

classifica

measured

R

and Holl

validity,

Example

F

This typ

definitio

must se

implem

behavio

Holland (1986) conducted a study to develop a criterion-referenced test of fundamental motor skills that is valid and reliable for elementary-aged students classified as learning disabled, educable mentally impaired or non-handicapped. Also, content validity was used by Voelker and Horvat (1974) to judge the validity of their instrument designed for determining the nature of elementary school children's environmental decisions. They judged content validity by using a panel of judges to validate items used in the instruments. The panel members, representing a wide range of perspectives on environmental education, gave their collective judgments on each item.

Content validity and concurrent validity were used by Colbert (1977) and Ruud (1976). In the study "The development and implementation of the Levels of Use Observational Inventory (LoUOI): An instrument to aid in the adoption of an innovation process," content validity was established by asking a panel of experts to inspect the LoUOI and the LoUOI classification, and to validate the items and the classification with the Levels of Use (LoU) Chart descriptions for each level. The instrument and classification were revised until total agreement was reached. Concurrent validity was measured by correlating the LoUOI ranks with the LoU interview using Sperman's Rho.

Ruud (1976), Voelker and Horvat (1974), Stallings (1978), Seidman et al., (1979), and Holland (1986) also used other types of validity, like construct validity, discriminate validity, convergent validity and face validity to judge the validity of their instruments. Examples of studies that used these types of validity follow.

Face validity was used in "The development of the contextual observation system." This type of validity was used carefully by the follow through models. The operational definitions are very straightforward. For example, to code "positive affect," an observer must see a broad smile or hear laughter. The code name reflects a concept whose implementation can be seen or heard (Stallings, 1978).

Discriminate and convergent validity were used in "Assessment of classroom behavior: A multiattribute, multisource approach to instrument development and

validation."

and peer an

(Seidman e

Reliabilit

Reliability

errors; that

instance, o

instrument.

error comp

scores as n

Rel

agreement

used as an

which the

been equa

1977).

Th

(Wittrock

1. When

a. B

b. T

c. C

b

2. On w

a.

validation." Here the effort to assess validity of the receptive components of the teacher, and peer and self behavior description forms, a multi-trait, multisource correlation matrix. (Seidman et al., 1979).

Reliability

Reliability is defined as an indication of the extent to which a measure contains variable errors; that is, errors that differ from individual to individual during any one measuring instance, or that varied from time to time for a given individual measured twice by the same instrument. Each measurement then consists of two components: a true component and an error component. Reliability is the ratio of the true-score variance to the variance in the scores as measured (Nachmias and Nachmias, 1976).

Reliability of observation refers to the consistency, stability and/or observer agreement. (Johnston and Pennypacker, 1980). The percent of observer agreement is also used as an indicator of observer reliability. Observer agreement indicates the degree to which the scores of observers who viewed certain events agree. Agreement not only has been equated with reliability, but has also come to represent the accuracy of data (Kazdin, 1977).

There are some questions related to estimating the reliability of an instrument (Wittrock, 1986).

1. When should observer agreement be measured?
    a. Before data collection.
    b. Training does not guarantee against observer skill deterioration across time.
    c. Calculation of the degree to which observer disagreement limits reliability should be done after the study.
2. On what kinds of data should observer agreement be calculated?
    a. Agreement should be computed on the same unit of behavior that will be used in the data analysis.

b.

3. W

a.

b

4. U

a

b

5. H

a

b

6. V

a

b

c

b. Agreement should be computed on sub-categories of behavior, as well as the larger measurement categories.

3. With whom should agreement be obtained?
   a. High inter-observer agreement is important but insufficient, because systematic misinterpretation can exist even with high agreement.
   b. Observers' scores should also be compared with a criterion. This is known as a criterion-related agreement.
4. Under what condition should agreement be calculated?
   a. Coding in the setting may be different from coding unambiguous samples in a laboratory or training session.
   b. Ways to heighten observer vigilance and maintain accountability should be considered.
5. How can agreement be measured?
   a. Intraclass correlation coefficients.

      It is useful after a study is completed, but impractical during or before, and it is highly affected by the variance between subjects.
   b. Simple percentage agreement.

      Drawbacks are that low frequencies in some categories and high frequencies in other may makes interpretations ambiguous, and it does not account for false inflation due to chance agreement.
6. Which agreement coefficient is appropriate?
   a. This is dependent upon the type of observation system, number of categories, type of data, unit of analysis, and purpose.
   b. If nominal comparison cannot be obtained, then marginal agreement methods should be used.
   c. If only a few categories and/or frequency distributions are unequal, then correction for chance agreement should be made.

d. The definition of "items" changes with systems. The probability of occurrence of the items must be considered.

Preparation or training of subjects who will judge the reliability of the instrument is an important condition that should be considered. For example, the Educational and Cultural Center (1969) conducted a study "Behavioral analysis instrument for teachers," to develop an instrument to describe teacher behaviors during classroom teaching, planning, evaluation and diagnosis. The authors emphasized the concept of preparation for the subjects who are going to judge the reliability of the instrument. They said that the outcomes can be considered reliable only when the instrument has been used by subjects prepared in its utilization.

The literature reveals that there is more than one method suitable for establishing the reliability of an instruments. Because each test or instrument has a different purpose, it is not necessary to use all of the methods of establishing reliability. Test-retest reliability is considered to be an important method to assess the reliability of tests and/or instruments because it is more generalizable than other methods of estimating reliability. Ruud (1976), Rose et al. (1973), Weston, Petosa and Pate (1997), Nachmias and Nachmias (1976), Pate (1995) and Colbert (1977), included examples of using test-retest methods to judge the reliability of their instruments.

The test-retest method corresponds most closely to the conceptual definition of reliability. A measuring instrument is administered to the same group of persons at two different times, and the correlation between the two sets of observations (scores) is computed. The obtained coefficient is the reliability estimate. With this method, error is defined as any thing that leads a person to get a different score on one measurement from what he or she obtained on another measurement.

The inter-rater method of judging reliability also has been considered to be appropriate, especially for instruments that will be used by observers. Different researchers with different purposes for developing their instruments have used the inter-

rater method to judge the reliability of their instruments. These include Rose et al. (1973), Stallings (1978), Weston, Petosa and Pate (1995), Colbert (1977), Voelker and Horvat (1974), Kazdin (1977) and Holland (1986). The inter-observer agreement refers to a situation in which the independent observation records of one observer are compared to those of a second person.

An example of a study that used both test-retest and inter-rater reliability was the study conducted by Weston, Petosa and Pate (1997). In this study titled "Validation of an instrument for measurement of physical activity in youth," both methods, were used. The inter-rater reliability coefficient for estimating relative energy expenditure (Previous Day Physical Activity Recall—PDPAR) for the entire day was 0.99 ($P<0.01$). The coefficient for estimating relative energy expenditure in play/recreation and exercise workout activities was 0.99 ($P<0.01$), and 1.0 ($P>0.01$), respectively. The test-retest reliability correlation coefficient for the PDPAR administered twice in one hour was 0.98 ($P<0.01$)

A second example of using inter-rater reliability was the study conducted by Stallings (1978). To judge the reliability of the development of the contextual observation system, inter-rater agreement in the classroom was used, which was assessed by using trainers as the criterion. Inter-rater agreement in classrooms was compared with scores on the criterion videotapes in a subsample of classrooms.

A third example was the study conducted by Colbert (1977), "The development and implementation of the levels of use observational inventory (LoUOI): An instrument to aid in the adoption of an innovation process." He used test-retest reliability and inter-rater reliability to judge the reliability of the system.

A fourth example was the study conducted by Holland (1986) to develop a criterion-referenced test of fundamental motor skills that is valid and reliable for elementary-aged students classified as learning disabled, educable mentally impaired or non-handicapped.

To estimate the reliability of his instrument he used inter-rater agreement reliability coefficient and the test-retest reliability coefficients.

Intra-rater reliability is one of the methods that can be used to judge the reliability of the instrument being developed in this study. In this regard, Kazdin (1977) stated the following facts.

1. Intra-observer agreement refers to the situation in which one observer makes an observation of events on one day and then comes back at a later point in time to observe the same events. In this situation the term "agreement" is closest in meaning to the term "reliability." The two observers start and end their observation at the same time, otherwise the agreement level will be lower. The time span between the two observation sessions should be at least one week, and preferably longer. The longer this time period, the less likely the observer will be able to remember their scores and thus replicate coding from the first observation session.
2. Independence for the intra-observer agreement procedure can be achieved by withholding observer access to the observation record of the first session when the second observation is made.

Other methods to judge reliability, are stated below.

1. Split-half reliability refers to a method that estimates reliability by treating each of two or more parts of a measuring instrument as a separate scale. For this method the instrument is separated into two sets, using the odd-numbered questions for one set and the even-numbered questions for the other. Each of the two sets of questions is treated separately and scored accordingly. The two sets are then correlated, and this is taken as an estimate of reliability. To correct the correlation coefficient obtained between the two halves, the formula known as the Sperman-Brown prophecy formula may be applied (Nachmias and Nachmias, 1976).
2. Parallel-forms reliability technique is a way of overcoming limitations inherent in the test-retest method. This technique requires two forms of a measuring instrument that

may be considered parallel. The two forms are then administered to a group of persons (or other objects), and the two sets of measures (scores) are correlated to obtain an estimate of reliability. With this technique, there is the problem of determining whether the two forms of an instrument are, in fact, parallel. Although statistical tests have been developed to determine whether the forms are parallel in terms of statistical measures, determination with respect to the content of the forms must be made on a judgmental basis (Nachmias, 1976).

3. Percentage of agreement. In a study conducted to develop a measure to evaluate language communication skills of young children, a pilot study to estimate test-retest reliability was conducted. In this study, 12 first grade children (not included in the sample for the following study), served as subjects. The classroom task was administered to the 12 children twice, one week apart. The children were randomly paired and randomly assigned to play the same role for both sessions. The percentage of agreement was calculated for each item. The mean percentage of agreement for a task was 89.3% with a range from 78.5% to 100%. In addition to the small pilot test-retest reliability study, the split-half and the parallel test methods were used to estimate the reliability (Rose et al., 1973).
4. Generalizability theory. In the development of an observation schedule for coding reading-related activities in secondary school classroom, reliability was estimated within the framework of generalizability theory. Classroom observation schedules are subject to more potential sources of unreliability than paper- pencil tests. These sources include infrequency or instability of teacher behavior, as well as rater inconsistency. While most observational studies report inter-rater consistency alone, instability of teacher behavior has been identified as an important source of instrument unreliability. Generalizability theory may be a more appropriate framework for conceptualizing and estimating the reliability of an observation schedule than traditional reliability theory because it allows the researcher to explicitly and simultaneously estimate more than one

source of score variation. Research on classroom observation schedule reliability indicates that generalizability can be difficult to achieve (Wolf and Greenwald, 1978).

5. Multiple methods. In a study titled "The development and field testing of an instrument to evaluate student personnel programs," more than one method was used to judge reliability. An odd-even reliability procedure and the Sperman-Brown prophecy formula were used to estimate the reliability of the instrument. The "yes" responses of 50 students, faculty and student personnel administrators to 36 questions were divided into two separate scores. The odd-even procedure resulted in a correlation coefficient of +.84 between the two halves of the instrument. The Sperman-Brown formula yielded a correlation coefficient of +.91 for the entire instrument, (McDavis, 1976).

   In another study, "The construction of an instrument to measure proportional reasoning ability of junior high pupils," more than one approach was used to judge the reliability of the instrument. As a criterion for reliability, it was expected that the same person (or a comparable person) taking the paper-pencil instrument should exhibit a comparable percentage of mastery. A classic one-form reliability measure was calculated. Individual pupil scores and the total number of correct responses were used. The criterion-referenced-nature of the testing and the scoring by category were used. The approach used in this study afforded a correction for the criterion level and the variance limitation of criterion-referenced testing. (2) Reliability was also measured on a test-retest basis and analyzed with the tetrachoric and Pearson correlation coefficients (Ruud, 1976).

Summary

Observational methodology is most appropriate when an argument can be made that some unique kind of information can be obtained by examining the behavior of a subject in a real life setting. The observation systems act as a mirror for clients and/or researchers to obtain feedback about classroom teaching behavior. Systematic observation has played a major role in the emergence of teaching/coaching behavior research as a bona fide area of

empirical study in our profession. There are specific steps involved in the process of systematic observation.

1. Deciding what to observe.
2. Developing definitions for the behaviors to be observed.
3. Selecting the most appropriate observation system.
4. Establishing observer reliability.
5. Making the actual observations.
6. Summarizing, analyzing and interpreting the data.

Observer error is one of the factors that influences the accuracy of the observational record. Errors can have their source in behavioral definitions, research design and problems of measurement. Another factor that influences data accuracy is the observers themselves. Some of the major obstacles that relate directly to observer error are observer drift, complexity of the observation system, observer expectancies/bias, observer reactivity and observer cheating.

Videotaping is an important tool for facilitating systematic observation. Videotaping feedback provides an objective recording of students' behaviors and classroom events, and therefore, is an important resource for assessment and instruction.

To develop a systematic observation instrument a number of questions need to be considered. These questions should be used to plan the development of an instrument.

1. What is to be measured?
2. How specific should the items be?
3. What context defines the demonstration of competence to be measured?
4. How often and at what points should subjects be measured?
5. Who will be involved in the measurement process?
6. Will the measurement be formative, summative, or both?
7. How much time will be required in adapting or developing the instrument?

Developing a valid and reliable systematic observation instrument requires several systematic steps. Because each instrument is created for a different purpose, different steps will be used in the development process. For this study, "Development of an instrument to assess the implementation fidelity of physical education lessons," the following steps should be appropriate for the development process:

1. Determine the purpose and objectives of the instrument.
2. Construct a table of test specifications for the instrument.
3. Create a pool of potential items.
4. Prepare instructions for using the instrument.
5. Use and refine the item pool.
6. Establish content validity of the instrument.
7. Establish reliability of the instrument.

It is important to create a clear statement of purpose and objectives to guide development of the instrument. Also important is the use of knowledgeable specialists to create the instruments purpose and objectives.

Constructing a table of test specifications is also an important step to provide the developer with a framework for the construction of the instrument and insure that appropriate content is included in the instrument. This table can be developed after determining the purpose and objectives of the instrument.

The item pool should be developed in accordance with the test specifications. The items then should be evaluated by content specialists according to specific criteria. Examples of criteria for this study include:

1. Match with the lesson components.
2. Consistency with the purpose of lesson component.
3. Consistency with the indicators of appropriate teacher implementation behavior.
4. Discreteness.
5. Measurability.

It is important to clearly communicate to users of the instrument its purpose and organization, as well as how to use and score the instrument. Instructions for using the instrument should include the following:

1. Introduction: Purpose of the instrument, overview of components, how to use the instrument and an example of how to score the instrument.
2. Rating scale: Description of the rating scale and the meaning of all levels of the scale.
3. Directions: Steps the users should follow to prepare for using the instrument.
4. Strategy: Description of ways of how to use the instrument efficiently. Main points for this instrument include:
    a. The instrument.
    b. Review of the MI-EPEC K-2 lessons.
    c. Review and use of the score sheet.
    d. Ways to conduct the assessment.

Any assessment instrument must be valid. The validity of the instrument can be established through more than one method. Appropriate methods include content validity, concurrent validity, construct validity, discriminate validity, convergent validity and face validity. The nature of the current study supports the use of face validity and content validity to describe test appropriateness. The instrument will be considered to have face validity when content experts rate the following aspects of the instrument as appropriate:

1. Operational definitions (clear and correct).
2. Instrument categories and sub-categories represent the lesson components.
3. The system for recording results of the observation correctly represents the levels of implementation behavior used by teachers implementing the lessons.

Content validity should answer the question: To what extent does the content of the instrument measure intended teacher implementation behavior? Content experts are most appropriate for making these judgments.

Reliability is defined as an indication of the extent to which a measure contains variable errors; that is, errors that differed from individual (or some other subject) to individual during any one measuring instance, and those that varied from time to time for a given individual on the same instrument. Percent of observer agreement is also used as an indicator of observer reliability. Observer agreement indicates the degree to which observers who viewed certain events agree in their recordings. Agreement not only has been equated with reliability, but has also come to represent the accuracy of data.

Preparing, or training, the observers also is an important issue in systematic observation. Proper training facilitates observer reliability in the use of an observation system. Based on the purpose of this investigation and the complexity of the observation system, various portions of the training may consume more time than others. The use of instructional principles such as successive approximations in teaching appropriate observation techniques is strongly recommended. The training process should include orientation to the system, learning the categories, using the coding form correctly, initial coding practice and live observation practice.

CHAPTER III

Methodology

Purpose of the Study

The purpose of this study was to develop an instrument to measure the degree to which physical education teachers implement K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project (MI-EPEC) as intended by lesson developers.

Development, as the term is used here, includes creating a valid and reliable instrument that will be appropriate to achieve the following three objectives:

1. To measure the degree to which teachers implement each of the nine components that appear in the lessons as they were intended by lesson developers.
2. To document the degree to which teachers implement each lesson component in accordance with the time allocated by lesson developers.
3. To provide a valid instrument for individuals who wish to document relationships that may exist between teacher implementation behavior emanating from use of MI-EPEC K-2 lessons and for example, inservice training devoted to use of the lessons or student achievement of lesson objectives.

Assumptions

Development of the instrument is based upon two basic assumptions:

1. The lessons of MI-EPEC are sufficiently described so that implementation fidelity can be measured.
2. Teachers who choose to do so are able to implement the lessons as intended by the developers.

Design of the Instrument

Review of the literature revealed several procedural steps to consider when developing an evaluation instrument. Each step and the method(s) used for its completion are described in the following paragraphs. The application of this literature to the development of the MI-EPEC Teacher Implementation Evaluation Instrument (TIEI) resulted in the use of seven steps as follows:

1. Determine the purpose and objectives of the instrument.
    1.1 Collect information related to the instrument's purpose and objectives from the Michigan Implementation Team (MIT).
    1.2 Create a prototype purpose statement and objectives.
    1.3 Obtain feedback from the MIT on the quality of the prototype statements.
    1.4 Finalize the instrument's statement of purpose and objectives.
2. Construct a table of test specifications.
    2.1 Construct a table of test specifications to provide a framework for item construction.
    2.2 Construct a pool of prototype items.
3. Review and revise the item pool.
4. Prepare instructions for using the instrument.
5. Use, refine and establish criterion scores for the items of the instrument.
    5.1 Use and refine the instrument.
    5.2 Use, refine the instrument, and establish criterion scores for the items.
6. Establish the content validity of the instrument.
7. Establish the reliability of the instrument.

Step 1: Determine the Purpose and Objectives of the Instrument

It is important to create a clear statement of purpose and objectives for the instrument being developed. Clarity of purpose and well articulated objectives are

necessary to guide the decisions at each step and sub-step of the instrument development process. The sub-steps used to create the statement of purpose and objectives for the instrument were as follows.

1.1 Obtain information from key staff members from the Michigan Exemplary Physical Education Curriculum Project (MI-EPEC). The purpose of this sub-step was to obtain information fundamental to developing prototype statements of purpose and objectives. Staff members selected were the project's curriculum and instruction specialist, evaluation specialist, and the project co-chair who fulfilled both curriculum and evaluation functions. This group was used for six of the instrument development steps and are referred to subsequently as the MI-EPEC Implementation Team (MIT). Information was obtained by responses of the MIT to the following four questions and one open ended item included to obtain other issues related to developing a purpose statement and objectives for the measurement of teacher implementation fidelity.

   a. Why is it important to measure implementation fidelity?
   b. What are you be willing to accept as evidence of high implementation fidelity?
   c. What kinds of teacher behaviors would be examples of low implementation fidelity?
   d. What variables do you believe will be associated with high and low implementation fidelity?
   e. Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity.

This sub-step was completed as follows:

1. Members of the MIT were asked to answer the four questions included on the form contained in Appendix A, Step 1.1: Implementation Survey.
2. All data obtained from each member was summarized as statements under each question.

3. All summarized statements were organized according to their commonalties (see Appendix B, Results of Step 1.1).

1.2 Create a prototype purpose statement and objectives.

Subsequent to summarizing the responses of the MIT, the information was divided into categories representing the statement's relevance to the instrument's purpose and/or objectives (see Appendix C, Analysis of the Statements for Their Potential Contribution to Creating a Statement of Purpose and/or Objectives for the Instrument). Drafts of the instrument's purpose statement and objectives were then created to represent the categorized information. The draft statements created were developed in accordance with several criteria. They were:

a. Clarity of language.

b. Representative of (or rooted in) MIT responses.

c. Clearly related to implementation fidelity.

d. Consistent with the total evaluation process of MI-EPEC.

The prototype purpose statement and objectives are included in part A (initial draft) of Appendix D, Development of the Statement of Purpose and Objectives for the Instrument.

1.3 Validate the appropriateness of the prototype statements.

The draft statements of purpose and objectives were returned to the MIT for review. The MIT members read and suggested changes necessary to maximize accuracy, appropriateness and clarity. Additions, alterations and deletions were requested along with supporting rationales. The suggestions obtained were incorporated into a revised draft consistent with the merit of their supporting rationales. This revised draft is included in part B (revised draft) of Appendix D.

1.4 Finalize the statement of purpose and objectives of the instrument.

The revised purpose statement and objectives of the instrument was returned to the MIT for review. The purpose and objectives were finalized when each member of the

MIT signed off on the last version. It was necessary to schedule meeting of the investigator and the MIT for the purpose of resolving remaining discrepancies in feedback. At this meeting all discrepancies were resolved by discussing the rationales of each member in terms of its importance to the suggested changes until all agreed about the final form of the purpose statement and objectives. The final statement of purpose and objectives are included in part C of Appendix D.

Step 2: Construct a Table of Test Specifications

Subsequent to final approval of the instrument's purpose and objectives, a table of test specifications was constructed. This table was designed to provide a framework for item construction. The table took the form of a matrix with the lesson components represented on one dimension and teacher implementation behaviors on the other. The teacher behavior side of the matrix was procedural in nature and led to specification of prototype items suitable for assessing teacher implementation of the lessons. The teacher behavior categories leading to the pool of potential items are as follows:

Column 1: Lesson components.

Column 2: Statement of the purpose for each lesson component.

Column 3: Identification of indicators acceptable to the MIT as evidence of full implementation of the intended purpose.

Column 4: A "place holder" for the creation of prototype item that represent the indicators identified in column 3.

The primary sources of information for completing each cell in the matrix and for developing the item prototypes were the criteria used to design, write and evaluate MI-EPEC lesson materials. These criteria were provided to the MIT as an aid to completing the test specification matrix. The matrix assured that all relevant components of the lessons would be evaluated. It also insured that appropriate teacher behaviors would be represented in the prototype items.

The MIT independently completed the "purpose" and "indicators" columns of the matrix. Although the MIT was intimately familiar with the structure and substance of the MI-EPEC lessons, they used the same instructional materials design lesson writing and evaluation criteria used by MI-EPEC lesson writers to create the lessons.

The instrument for collecting test specification information from the MIT is included in Appendix E, Data Collection Forms for Step 2: Creating a Table of Test Specifications. The MIT responses were compiled, edited and then returned to each MIT member for review along with their original submission. A meeting was then held to resolve discrepancies and a new consensus version was created, edited and finalized for use in the item generation step.

The primary source of information used to create prototype items was the information included in the "indicators" column of the matrix. Prototype items were created for each lesson component represented in the test specification matrix.

Step 3: Review and Revise the Item Pool

The prototype items were evaluated by the MIT for appropriateness of content and format. The evaluation criteria used were as follows:

1. Clarity.
2. Consistency of the item with the indicators of appropriate implementation behavior.
3. Discreteness of the behavioral description.
4. Measurability, including appropriateness of the response format.
5. Consistency of the item with the lesson component.
6. Lack of bias.
7. Others.

The item pool was revised in accordance with the suggested changes. Items were eliminated, added and altered. The strength of the rationale associated with the proposed change was the basis for the changes made. The items were then resubmitted to the MIT

for final approval. Some of the changes created additional discrepancies and some need additional work to gain MIT approval. These changes were resolved in face-to-face meetings devoted to establishing consensus on an item-by-item basis.

Step 4: Prepare Instructions for Using the Instrument

It is important to clearly communicate the instrument's purpose and organization, as well as how to use and rate the items. Information contained in the purpose and objectives statements, the table of test specifications and items included in the refined item pool provided the information necessary to format the instrument and write instructions for its use.

The format of the instrument and instructions for its use were reviewed by the MIT. The MIT was asked to read the instructions and react to their appropriateness in describing the purpose of the instrument, how it should be used, and its format. Again MIT members were asked to provide a rationale for substantive changes recommended. Revisions reflecting this feedback were then made by the investigator.

The instructions for using the instrument included the following categories:

1. Introduction
2. Purpose
3. Overview
4. Response format
5. Measurement procedures

Step 5: Use, Refine and Establish Criterion Scores for the Items of the Instrument

5.1 Use and refine the instrument

Subsequent to revising the instructions, the MIT pilot tested and revised the instrument. This pilot test involved using the instrument to rate implementation

fidelity of a videotape of one teacher teaching a 15 minute MI-EPEC lesson. After the MIT rated the lesson, a meeting was held to identify and discuss the difficulties faced while using the instrument and how they could be avoided. Again revisions to the instrument were made based on the problems encountered during the rating process. Sub-step 5.1 also identified items that were redundant, unclear or inconsistent with the criteria specified. It also provided information helpful to improving the instructions for use and the instrument's format. Changes were made in accordance with the strength of their supporting rationales and incorporated into a revised instrument.

5.2 Use, refine and establish criterion scores for the items

The instrument was used a second time by the MIT to rate implementation fidelity of another teacher teaching a 15 minute segment of a MI-EPEC lesson. The procedures used in 5.2 were the same as those used in 5.1. Again the suggested changes obtained were integrated into the instrument and a criterion score for each item was established. This criterion score was used for judging the reliability of the instrument. When the Student Rating Team (SRT) used the instrument (step 7) to rate this same teacher, their scores were related to the criterion scores.

Step 6: Establish the Content Validity of the Instrument

This step was designed to describe the content validity of the instrument. Content validity was used to answer the following question: To what extent does the content of the instrument measure intended MI-EPEC teacher implementation behavior? The answer to this question was based on the professional judgment of the MIT members. Two primary questions guided the collection of data necessary to complete this part of the study.

1. Does the instrument represent all components of the MI-EPEC lessons?
2. Do the items appropriately represent teacher implementation behaviors included in the MI-EPEC K-2 lessons.

To obtain evidence of content validity, each member of the MIT completed the form designed for this step (see Appendix F, Step 6: Assessing Content Validity of the Instrument) and circled the responses that represented his expert opinion.

Step 7: Establish the Reliability of the Instrument

The inter-rater reliability of the instrument was established by using the data obtained from MIT and SRT members who used the instrument. Rating responses and criterion scores obtained from the MIT and SRT were used to calculate the inter rater reliability. The correlation coefficients between all possible pairs of raters were calculated using the Pearson correlation procedure contained in the Statistical Package for the Social Sciences (SPSS). Inter rater reliability was calculated for the components of the instrument and for the instrument as a whole.

For estimating the reliability of the instrument as a whole (only continuous responses were used) the following mean correlations were obtained between all possible pairs of raters

1. Mean correlation between the SRT and criterion.
2. Mean inter-rater reliability among the SRT members.
3. Mean inter-rater reliability among the MIT members.

The same mean correlations were then calculated for the categorical responses "Not Applicable" and "Not Ratable" (NA & NR). Again correlations were established between all possible pairs. This correlation was obtained by coding each continuous response as a "1" and each categorical response as a "2" then calculating the agreement between pairs. The first two components (equipment/materials and gym setup) were combined because each component only has 4 items and their content was similar.

Item responses by the MIT and SRT were also analyzed by response categories and a degree of agreement coefficient was calculated. The agreement coefficient describes the degree to which ratings were encompassed by adjacent rating categories.

Data Collection Procedures

The chairperson of the guidance committee and co-director of the MI-EPEC project was consulted to identify the subjects for the study. The MIT members from Michigan State University were contacted by the investigator to obtain their agreement to participate. The SRT members from the University of Michigan were contacted by Dr. Charles T. Kuntzleman, co-director of the MI-EPEC project and faculty member at the University of Michigan.

All data related to the seven steps used to develop the instrument were collected between April 30 and July 2, 1998. Data were collected in accordance with the steps described earlier. Data collection for some steps, required follow up meetings to discuss and obtain consensus. Seven meetings averaging two hours were needed to complete the steps. For each step, the subjects of the study were contacted by phone or E-mail to provide them with the necessary information and/or data collection forms.

To obtain the data for step 7, "establishing reliability of the instrument," a one day inservice training session was scheduled for the SRT. This session was conducted in accordance with the agenda included in Appendix G, Inservice Agenda, by Dr. Paul Vogel, who volunteered to help the investigator.

Data Analysis

Based on the type and/or form of information (qualitative, quantitative) obtained from each step, it was analyzed in the following manner.

Step 1: Determine the purpose and objectives of the instrument.

1.1 Survey selected MI-EPEC staff members (MIT)

a. Data obtained from the MIT were in the form of responses to four questions and one open-ended response (see Appendix A).

b. Data from each subject for each question were analyzed by reading the responses, categorizing "like" information, and converting it to summary statements.

1.2 Create a prototype statement of purpose and objectives.

a. All statements obtained in sub-step 1.1 were checked as "P" (statement of purpose) or "O" (statements of objectives) in accordance with their potential contribution to the development of the purpose statement or statement of objectives.

b. This information was then used to guide development of the prototype purpose and objective statements.

1.3 Review the prototype statements of purpose and objectives.

a. Results obtained from the MIT were in the form of qualitative data including suggested changes, additions, deletions, alterations, etc. and the rationales for the changes.

b. All of the suggested changes were integrated into the statements of purpose and/or objectives.

1.4 Finalize the statements of purpose and objectives.

a. All MIT members reviewed the revised statements of purpose and objectives.

b. Members then signed off on these statements designating their final approval.

Step 2: Construct a table of test specifications.

a. All data obtained from the MIT were organized under each component included in the MI-EPEC lessons.

b. The purpose statements and indicators obtained for each component were converted to prototype items.

Step 3: Review and revise the item pool.

a. Changes suggested by members of the MIT were considered for incorporation based on the strength of the rationale supporting the suggested change and then integrated into the instrument.

Step 4: Prepare instructions for using the instrument.

a. Instructions were prepared to describe how to use the instrument.

b. Changes suggested by members of the MIT were considered for incorporation based on the strength of the supporting rationale and then integrated into the instrument.

Step 5: Use, refine and establish criterion scores for the items of the instrument.

5.1 Use and refine the instrument

a. All rating scores obtained from the use of the instrument were recorded.

b. Variances in scores were resolved by meeting, discussing the variances, and their rationales and then resolving differences based on the strongest rationales.

5.2 Use, refine the instrument and establish criterion scores for the items

a. All rating scores obtained from the use of the instrument were recorded.

b. Variances in scores were resolved by meeting and discussing the variances, considering the rationales for the varied ratings and then resolving differences based on the strength of their supporting rationales.

The resolved differences thereby established the criterion scores for each item.

Step 6: Establish the content validity of the instrument

a. Responses obtained from the MIT in regard to the representation of the instrument's components to the lesson's components and appropriateness of

the items contained in the instrument were in the form of "Yes" or "No" responses.

b. Yes and No responses were summarized.

c. Judgments regarding the appropriateness of items were made on the basis of these responses.

Step 7: Establish the reliability of the instrument.

a. The Pearson correlation coefficient was used to describe the inter-rater reliability.

b. Descriptive statistics (agreement coefficient) was used to describe rater agreement.

Training Observers

Carpenter (1977), Stallings (1978), and Smith, Smoll and Hunt (1977) stated that proper use of the observation system requires intensive observer training. Unless independent observers can agree on how a particular behavior is to be categorized, the system cannot be scientifically useful. Thus the major goal of any training program should be to establish high inter-rater reliability. Smith, Smoll and Hunt (1977) pointed out that to achieve high inter-rater reliability inservice should explain the observation system, provide extensive practice, use written tests in which the trainees are required to define the system and score behavioral examples and show a high degree of expertise in the use of the system before data they generate is used for research purposes.

CHAPTER IV

Results of the Study

The purpose of the study was to develop an instrument suitable for measuring the degree to which teachers implement MI-EPEC K-2 lessons as written by the lesson developers. The purpose of this chapter is to present the results obtained from each of the seven steps used to develop this instrument. Results presented in this chapter were collected from seven subjects; three of them were staff members of the Michigan Exemplary Physical Education Curriculum Project, referred to here as the MI-EPEC Implementation Team (MIT). The other four were kinesiology graduate students enrolled at the University of Michigan who were employed in a related project funded to test student achievement of selected high priority physical education objectives using videotape vignettes. They will be referred to here as the Student Rating Team (SRT). The MIT was used for steps 1-6 of the development process and the SRT was used for step 7. The data obtained for each step is presented here in the order of the seven steps described in Chapter 3.

Step 1: Determine the Purpose and Objectives of the Instrument

Overview

The purpose of step 1 was to create a clear statement of purpose and objectives for the instrument. A statement of purpose and objectives were viewed as important guides to the rest of the development process.

Four sub-steps were used to complete step 1. They were: 1.1) collect information related to the instrument's purpose and objectives from the MIT; 1.2) create a prototype purpose statement and objectives; 1.3) obtain feedback from the MIT on the quality of the prototype statements; and 1.4) finalize the instrument's statement of purpose and objectives.

Results of step 1.1 and 1.2

The data obtained from the MIT consisted of responses to the following five questions:

1. Why is it important to measure implementation fidelity?
2. What are you willing to accept as evidence of high implementation fidelity?
3. What kinds of teacher behaviors would be examples of low implementation fidelity?
4. What variables do you believe will be associated with high and low implementation fidelity?
5. Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity.

The instrument used to collect data for sub step 1.1 is included in Appendix A. There were 14 responses to question number 1; two of these responses were obtained from two members of the MIT. There were nine responses to question number 2; three of these responses were obtained from two members of the MIT. There were 24 responses to question number 3; five of these responses were obtained from two members of the MIT. There were 23 responses to question number 4; two of these responses were obtained from two members of the MIT and one response was obtained from three MIT members. There was one response to question number 5. In all, the MIT members provided 71 responses to the five questions posed.

A summary of the MIT responses to each question revealed the following commonalities. Each summary response "statement" is followed by reference to the original MIT statement included in Appendix B.

Question one: Why is it important to measure implementation fidelity?

Knowledge of implementation fidelity:

1. Is critical to systematically improve the instruction in the next edition of the lesson (A1, A2, A3, A4, A6, A7, A9, A12 and A13).
2. It will reveal the degree to which teachers will implement lessons written by others (A5). (This response was obtained from two members.)

3. It will provide a direct measure of the effectiveness of the inservice education devoted to implementation of MI-EPEC K-2 lessons (A8 and A10).
4. It will provide a description of instruction as it actually occurred (A11).
5. It documents the degree to which instruction was implemented as written by lesson developers (A14).

Question two: What are you willing to accept as evidence of high implementation fidelity?

High implementation fidelity will be achieved when:

1. Teacher behavior matches the teaching behaviors articulated in the lesson (B1, B2, B4, B5, B6, B7, B8 and B9).
2. Student behaviors represent expectations drawn from the lesson description (B3).

Question three: What kinds of teacher behaviors would be examples of low implementation fidelity?

Low implementation fidelity occurs where:

1. Teacher implementation behavior does not match the teaching behaviors articulated in the lesson (C1, C2, C3, C5, C6, C7, C8, C11, C12, C14, C15, C16, C17, C18, C19, C20 and C21).
2. Students are engaged in behaviors/activities not included in the lesson description (C4, C22, C23 and C5).
3. Teacher spends time managing interruptions (C9, C10 and C24).
4. Teacher engages only a portion of the students, rather than all students, in the activities of the lesson (C13).

Question four: What variables do you believe will be associated with high and low implementation fidelity?

The following variables were identified by the MIT:

1. Clarity of the lesson's written description (D1).
2. Inservice preparation of the teachers (D2, D5 and D16).
3. Length of the lesson (D3).
4. Teacher commitment to faithfully teach MI-EPEC lesson content (D4 and D13).
5. Teacher comfort with methods described in the lesson (D6 and D14).
6. Class size (D7).
7. Expectations teachers hold for student learning (D8, D11 and D12).
8. Philosophical orientation of the teachers (D9).
9. Nature of the students (D10, D20, and D22)
10. Incentives associated with the provision of effective instruction (D15 and D23).
11. Load (D17 and D18).
12. Equipment/facilities (D19).
13. Experience of the teacher (D21).

Open-ended question: Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity.

The following response was the only one received for this question.

> Assessment of implementation of MI-EPEC lessons is different than the assessment of effective instruction.

All of the MIT statements included in Appendix B were also analyzed for their potential contribution to creating a statement of purpose and/or objectives for the instrument (see Appendix C). A check in the "P" or "O" columns identifies the statement as appropriate to consider in framing Purpose (P) and/or Objective (O) statements for the instrument. Where a statement was appropriate for both purposes both columns are checked. Where it was judged as inappropriate for either use both columns were left blank.

Results of assessing the statements of use revealed that 39 statements (out of 71) were checked for both purpose and objectives; one statement was checked just for purpose; 12 statements were checked just for objectives; and 19 statements were not checked for either the purpose or the objectives. Results of this process are included in Appendix C.

All checked statements in the "P" or "O" columns were used to create a prototype purpose statement and objectives for guiding development of the instrument. This process resulted in the initial draft statement of purpose and objectives (see part A of Appendix D).

Results of sub-step 1.3 and 1.4

All members of the MIT completed step 1.3, "review prototype statement of purpose and objectives" to validate the appropriateness of the prototype statements created in step 1.2. Results obtained were in the form of feedback (additions, deletions and alterations) to the drafted prototype statements.

The content of revisions suggested by MIT members to the statements of purpose and objectives were similar in substance but quite different in form. One of the members suggested four objectives, another five objectives, and the third suggested seven objectives. Their reviews of the statement of purpose and objectives are included in part B of Appendix D. The variation in suggestions was sufficient to add step 1.4 to the development process to finalize the statement of purpose and objectives.

Finalizing the purpose and objectives statements was accomplished by the members of the MIT meeting with the investigator to review and discuss the rationales for the alternative statements of purpose and objectives. Agreement on one statement of purpose and three objectives was accomplished during this meeting. The final statement of purpose and objectives approved by MIT members are included in part C of Appendix D.

Step 2: Construct a Table of Test Specifications

Overview

Two sub-steps were used to complete step 2. Sub-step 2.1 was devoted to constructing a table of test specifications to provide a framework for item construction while sub-step 2.2 was devoted to creating the item pool. A narrative form was used in sub-step 2.1 to obtain the test specification data (see Appendix E). Completion of this form required multiple responses by the MIT to the nine components included in MI-EPEC K-2 lessons. Under each component, the MIT member was asked to state a purpose(s) for the component and then to write indicators he would accept as evidence of appropriate teacher implementation of the component. Sub-step 2.2 was devoted to converting the indicators of sub-step 2.1 to items.

Results

Results obtained from the completion of the data collection form (see Appendix E) by the MIT members included purpose statements and indicators for each lesson component. Their responses are included in Appendix H, Purpose Statements and Indicators Obtained from the MIT for Each Lesson Component. The results obtained for each lesson component are summarized in the following table.

Table 1 - Number of purpose statements and indicators for each lesson component that were obtained from MIT members.

#	Components	# of Purposes Obtained	# of Indicators Obtained
1.	Equipment and materials	3	13
2.	Gym setup	5	14
3.	Preparation of students	4	19
4.	Explanation/demonstration	2	28
5.	Transition	2	10
6.	Practice	6	25
7.	Lesson review	2	13
8.	Lesson summary	3	13
9.	Homework	2	10
10.	General	1	3

Review of the indicators obtained in sub-step 2.1, in conjunction with the purpose statements obtained in the same sub-step, provided strong guidance for creating prototype items. Each indicator obtained was assessed for potential contribution to the pool of items based on its content. Some indicators needed only minor revisions to form an item. Others were complex in nature and provided the basis for several items. Formation of items was guided by four primary criteria: measurability, discreteness, clarity, and lack of redundancy. Accordingly, each indicator was worded so that all indicators were represented in the item pool. Items were then reviewed for redundancy. All redundant items were eliminated and items that were ambiguous were reworded to clearly communicate the underlying indicators. This process resulted in the creation of 115 items from the 148 indicators obtained from MIT members.

Step 3: Review and Revise the Item Pool

Overview

The purpose of this step was to review and revise the prototype items that were created from the indicators obtained in sub-step 2.2. Members of the MIT were asked to rate the appropriateness and/or inappropriateness of each item and its response format.

Results

Evaluation by members of the MIT of the appropriateness and/or inappropriateness of the items and response format resulted in items being eliminated, added, changed and/or included as written. Of the 115 items in the item pool, 32 (28%) were judged appropriate as written by all members of the MIT. Another 38 items (33%) were rated as appropriate by two of the three members of the MIT and 36 items (31%) were rated as appropriate by one member of the MIT. All members of the MIT identified eight items (7%) as inappropriate. Two of the three members identified another 36 items (31%) as inappropriate and at least one member identified 38 items (33%) as inappropriate. There was no answer to one item.

Variances in the appropriateness of the prototype items, as judged by members of the MIT, were resolved during three meetings, each of which covered the items from three lesson components. In these meetings, disagreements in the independent ratings with the rationales for the item-by-item variances were discussed. Items were altered, deleted and, in some cases, created to resolve the problems identified based on the strength of the rationales associated with the proposed solutions.

A summary of the revisions made during and subsequent to the three meetings are included in Table 2.

Table 2 - Summary of revisions made to the original item pool.

Lesson Component	Original Items	Items Deleted	Items Added	Items Altered	Final Items
1. Equipment/materials	9	5	-	4	4
2. Gym setup	12	10	2	2	4
3. Preparation of students	15	6	-	8	9
4. Explanation/demonstration	20	8	-	11	12
5. Transition	10	7	-	2	3
6. Practice	18	8	2	7	12
7. Review	10	6	3	4	7
8. Summary	11	7	3	4	7
9. Homework	7	5	4	1	6
10. General	3	3	-	-	-
Total items	115	65	14	44	64

Items that were identified as appropriate or inappropriate by all members of the MIT were not discussed in the meetings held to resolve item appropriateness. They were included as written when they were rated appropriate or deleted when rated inappropriate. All discussion centered on items where agreement was not unanimous.

The Strongly Agree/Strongly Disagree response format was approved by the MIT. Special emphasis was given, however, to the importance of clearly defining each response option in the instructions for use of the instrument.

Step 4: Prepare Instructions for Using the Instrument

Overview

The purpose of this step was to develop instructions for using the prototype instrument. Draft instructions were developed which included the following parts: introduction, purpose of the instrument, overview of MI-EPEC lesson components, response format, and procedures for conducting the measurement process. The initial draft and directions for its review are included in Appendix I, Step 4: Prepare Instructions for Using the Instrument. The MIT members were asked to judge the appropriateness of the instructions and suggest how they could be improved.

Results

Information obtained from the MIT members was qualitative in nature. The feedback obtained focused on additions, deletions and/or alterations of various portions of the instructions as follows.

For the introduction and purpose of the instrument, there were few changes suggested. An example that was typical of the kind of suggestions obtained was the suggestion to change the statement, "the degree to which teachers use" to "the degree to which teachers implement."

Changes suggested for improvements to the "overview of the lesson components" portion of the instructions included adding a clear description of the nine lesson components of each MI-EPEC lesson and how the items serve as indicators of desired teacher behavior. The overview was revised in accordance with the suggestion as follows:

> Implementation of the MI-EPEC lessons as they are written requires nine areas of teacher behavior. Each of these areas, referred to as lesson components (equipment/materials, gym setup, preparation of the students, explanation/demonstration, transitions, practice, review, summary, homework), is quantified by reviewer responses to a number of assessment items. Each item is an indicator of how fully a lesson component was implemented. Reviewer responses to the items are measured using a modified Likert scale.

All changes in the introduction were made in accordance with the strength of the rationale associated with the suggestion.

The data obtained from MIT members also suggested changes in the item response format. Changes suggested included revising the order of the response options to begin with "full implementation" and end with "not implemented." There were also significant changes to the definitions of the scale levels. The nature of these changes are revealed by comparing the descriptions in Appendix I with those in Appendix J, Teacher Implementartion Evaluation Instrument (TIEI).

One additional response option was added for users of the instrument. A "Not Applicable" (NA) response was needed for instances when an item is present but the written lesson does not call for the teacher behavior represented by that item.

There were also changes made to the measurement procedures section. Responses revealed that one of the procedures used to familiarize testers with the videotape player omitted "how to use the headphones." There were several editorial changes that improved clarity and some organizational improvements that reduced the length of this section while maintaining and/or improving clarity. There were no changes suggested for the "evaluator information sheet."

Step 5: Use, Refine and Establish Criterion Scores for the Items of the Instrument

5.1 Use and Refine the Instrument

Overview

This step was designed to pilot test the second draft of the instrument (Appendix K, Second Edition of Instrument Obtained as a Result of Completion of Steps 1-4). The pilot test involved each MIT member using the instrument to rate a videotape of the implementation behavior of a teacher teaching a 15 minute segment of a MI-EPEC lesson. The intent of this step was to:

1. Determine if the items in all components could be used by experts in an authentic assessment setting.
2. Identify difficulties that caused variance in the ratings of individual MIT members.
3. Create a criterion score for each item that could be used during the inservice training of SRT members.

Results

Results obtained from the pilot test achieved the intents specified above. The data revealed that there were agreements and disagreements in the MIT ratings of teacher implementation behavior and that several rating difficulties were present. To understand and resolve the difficulties, a meeting was conducted to discuss each rating discrepancy that occurred and the difficulties that emerged from the discrepancies obtained.

During the post rating meeting, items rated differently by members of the MIT were reviewed with a focus on the independent rationales stated for the ratings given. Discussion continued until the MIT members resolved the problem. When necessary, the videotaped lesson was reviewed as much as necessary to obtain resolution. Resolution also resulted in establishment of a criterion score for the item.

The recommended item changes are summarized below by lesson component.

Equipment/materials: There were a few editorial changes necessary to use the items included in this component. Most can be described as substituting one word for another. The number of items remained the same and there were no additions or deletions.

Gym setup: There was one suggested change which was to add the parenthetical phrase (per student) to item number 1 in Appendix L, Third Edition of Instrument Obtained as a Result of Completion of Step 5.1.

Preparation of students: Items 2 and 5 of Appendix K were reworded and one item was added. The revised items became items number 2, 5 and 6 of Appendix L.

Explanation/demonstration: One item was deleted (number 1) and a few editorial changes were made.

Transition: There was an extensive discussion resulting in agreement among the three members of the MIT that this component should not be a separate component of the instrument. Rather, the items included in this component should be replicated in the beginning of other components where MI-EPEC lessons typically include transitions.

Practice: There were two items slightly altered.

Review: There was one item altered.

Summary: There were no changes.

Homework: There were two items altered.

Results obtained from the MIT members engaged in the post rating meeting also included recommendations for changes in the instrument's instructions, rating scale, and organization of the instrument's items. These changes are briefly described below.

The instructions for the instrument were usable, however, there were a few editorial changes, such as replace the statement "use K-2 physical education lessons" by the statement "implement K-2 physical education lessons." There were also a few grammatical changes.

To properly rate the videotape, members of the MIT found they needed another response option. A "Not Ratable" (NR) response was added to provide an appropriate response for instances where the videotape record was insufficient to judge the degree of implementation.

The MIT members also agreed that only the SA-SD portion of the scale (contrary to the implementation descriptions typically used) should be used to rate the last two items of the following components: 1) Preparation of students; 2) explanation/demonstration; 3) practice; 4) review; 5) summary; and 6) homework. The two items are: 1) The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component; and 2) The teacher managed unanticipated

events (distractions) by quickly refocusing attention to the intent of this component of the lesson.

Since these items describe teaching behaviors that could affect student achievement but are not teaching behaviors that are part of the lesson narrative, the "implementation description" portion of the scale was considered inappropriate.

These changes suggested by the MIT were incorporated into a revised instrument. This third edition instrument is included in Appendix L.

Step 5.2: Use, Refine the Instrument, and Establish Criterion Scores for the Items

Overview

After the instrument was revised in step 5.1, it was used a second time by the MIT. The purpose of step 5.2 was to repeat the "use" based revision of the instrument by rating teacher implementation behavior on another 15 minute MI-EPEC lesson segment. The intent of this step was to:

1. Identify problems of use associated with another teacher and different lesson content.
2. Resolve the problems encountered.
3. Create a criterion score for each item so that the accuracy of ratings of other users could be determined by comparing the degree to which their scores matched the criterion scores.

Results

Results obtained from the MIT members who used the instrument a second time revealed variances in the ratings. In this round all three MIT members agreed on 44 of the 76 items (58%) included in the revised instrument.

To describe the amount of agreement across both continuous (1-5) and categorical (NA, NR) response options, the MIT scores were analyzed by response category. These data are summarized in Table 3. The table located in Appendix M (Degree of Agreement

Among Members of the MIT) shows the degree of agreement across all items (column 1) for the continuous items (column headings equal rating responses 1-5) and the categorical items (columns headed NA and NR). It also shows the criterion score established by the MIT. The degree of agreement on continuous items only and on all responses are included in the last two columns.

Of particular importance to correct interpretation of the table in Appendix M are the definitions of the degree of agreement associated with the numbers 1-6 contained in the last two columns of the table. As indicated, a "1" here indicates that only one response option was necessary to contain the ratings of all raters. A "1" therefore would equal perfect agreement on that item. A "2" in this column would indicate that two adjacent response categories were necessary to contain the ratings of all raters. The term "adjacent" is particularly meaningful in that the "closeness" of the ratings is important as well as "agreement" in the ratings. For example if half of the ratings were rated 4 and half were rated 5 on the continuous scale, the agreement value would be "2," indicating that 2 adjacent response categories contained all of the ratings. If the half of continuous response ratings were 1 and half were 5, the agreement value would be "5" indicating that five adjacent response categories were necessary to contain the ratings obtained. Because the two categorical response options, NA and NR, would be scored differently if their order were reversed in the table, they were considered one category for this analysis and one response category would be added to the number of response categories needed to contain the scores.

A coefficient of agreement was created to portray the overall agreement among members of the MIT. The coefficient of agreement is the mean of the agreement values (1-5) portrayed in the last two columns of the table located in Appendix M. The data are summarized for the MIT in Table 3.

Table 3 - Summary of the degree of agreement among the three members of the MIT on continuous and categorical data.

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	44	(58)	**Agreement Coefficient = 1.657**
	2	20	(26)	
	3	6	(8)	
	4	6	(8)	
	5	0	(0)	
	6	0	(0)	
	Total	76	100%	

For all items in the instrument the degree of agreement across all responses was as follows. For MIT ratings, results show that 44 items (57.8%) had perfect agreement; 20 items (23%) had all ratings contained in two adjacent response categories; six items (7.9%) had all ratings contained in three adjacent response categories; six items (7.9%) had all ratings contained in four adjacent response categories; and there were no items that needed five or six adjacent response categories to contain all of the ratings.

Thirty-two items received different ratings by at least one MIT member. To help understand the nature of the variances portrayed in Table 3, a meeting was held to discuss why individual MIT members rated individual items differently and to make refinements necessary to obtain consensus on a criterion score for each item. The same procedures used in step 5.1 were used for step 5.2. Each member shared their item score and their rationale for the rating. When necessary, the videotaped lesson was reviewed until all members agreed on a criterion score.

The discussion related to resolving rating variances and creating criterion scores resulted in additional refinements to the instrument. Some items were changed and transition items were moved from the beginning of each component to the end of each component to better connect the items to where transitions normally occur in the MI-EPEC lessons. A note was also added to explain when to use the three transition items.

Five items were altered according to changes suggested by the MIT to improve agreement. The changes are described below:

1. The original item, "The number of teaching stations described in the lesson component were used" was revised to read, "The number of teaching stations (areas were students participate interdependently on a learning task) described in the lesson component were used."
2. The original item, "The kind of teaching stations specified in the lesson component (distances, target sizes, available space . . .) were used" was revised to read, "The kind of teaching stations specified in the lesson component were used."
3. The original item, "If applicable (when called for in the lesson) students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated" was revised to read, "Students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated as written."
4. The original item, "The distribution, positioning, use and retrieval of equipment was accomplished as described in the lesson" was revised to read, "The distribution, positioning, use and retrieval of equipment/materials was accomplished as described in the lesson."
5. The original item, "When called for by the lesson, the teacher altered the lesson objective for a student(s) they achieved the criteria stated in the practice activity" was revised to read, "When called for by the lesson, the teacher altered the lesson objective for a student(s) who achieved the criteria stated in the practice activity."

The final draft of the instrument, consisting of 76 items covering eight lesson components, is included in Appendix J.

The MIT responses were analyzed in two additional ways for inter rater reliability. For the items where no categorical responses were used, the mean inter-rater reliability was calculated to be .77 (SPSS, Pearson product moment across all pairs of raters). For the items that received categorical responses from at least one member of the MIT, mean inter-rater reliability was .90. This correlation was established by coding continuous responses as "1" and categorical responses as "2" and then running the SPSS, Pearson correlation.

Descriptive responses for the components of the instrument

To portray the amount of agreement among raters across both continuous and categorical responses for the MIT on each component of the instrument, the ratings were analyzed by response category within each component of the instrument. Data are included in tables 4-11 and summarized in Table 12.

Again, a "coefficient of agreement" (COA) was created to portray the overall agreement in ratings among members of the MIT on each of the eight components. The degree of agreement across all responses were as follows.

1. For component number 1, "Equipment/materials."

   Results show that three items (75%) had perfect agreement; one item (25%) had all ratings contained in four adjacent response categories; and there were no items that needed two, three, five or six adjacent response categories to contain all of the ratings.

Table 4 - Agreement coefficient among the three members of the MIT for component number 1, "Equipment/materials."

Raters	Agreement Values	Number of items	Percent of Items	
MIT	1	3	75.	**Agreement Coefficient = 1.75**
	2	0	0	
	3	0	0	
	4	1	25.	
	5	0	0	
	6	0	0	
	Total	4	100%	

2. For component number 2, "Gym setup."

   Results show that one item (25%) had perfect agreement; three items (75%) had all ratings contained in two adjacent response categories; and there were no items that needed three, four, five or six adjacent response categories to contain all of the ratings.

Table 5 - Agreement coefficient among the three members of the MIT for component number 2, "Gym setup."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	1	25.	**Agreement Coefficient = 1.75**
	2	3	75.	
	3	0	0	
	4	0	0	
	5	0	0	
	6	0	0	
	Total	4	100%	

3. For component number 3, "Preparation of students."

Results show that three items (30%) had perfect agreement; four items (40%) had all ratings contained in two adjacent response categories; three items (30%) had all ratings contained in three adjacent response categories; and there were no items that needed four, five or six adjacent response categories to contain all of the ratings.

Table 6 - Agreement coefficient among the three members of the MIT for component number 3, "Preparation of students."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	3	30.	**Agreement Coefficient = 2.00**
	2	4	40.	
	3	3	30.	
	4	0	0	
	5	0	0	
	6	0	0	
	Total	10	100%	

4. For component number 4, "Explanation/demonstration."

Results show that nine items (64%) had perfect agreement; five items (38%) had all ratings contained in two adjacent response categories; and there were no items that needed three, four, five or six adjacent response categories to contain all of the ratings.

Table 7 - Agreement coefficient among the three members of the MIT for component number 4, "Explanation/demonstration."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	9	64.28	**Agreement Coefficient = 1.35**
	2	5	35.7	
	3	0	0	
	4	0	0	
	5	0	0	
	6	0	0	
	Total	14	100%	

5. For component number 5, "Practice."

Results show that 11 items (73%) had perfect agreement; two items (13%) had all ratings contained in two adjacent response categories; one item (7%) had all ratings contained in three adjacent response categories; one item (7%) had all ratings contained in four adjacent response categories; and there were no items that needed five or six adjacent response categories to contain all of the ratings.

Table 8 - Agreement coefficient among the three members of the MIT for component number 5, "Practice."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	11	73.3	**Agreement Coefficient = 1.46**
	2	2	13.3	
	3	1	6.66	
	4	1	6.66	
	5	0	0	
	6	0	0	
	Total	15	100%	

6. For component number 6, "Review."

Results show that five items (50%) had perfect agreement; three items (30%) had all ratings contained in two adjacent response categories; two items (20%) had all ratings contained in four adjacent response categories; and there were no items that needed three, five or six adjacent response categories to contain all of the ratings.

Table 9 - Agreement coefficient among the three members of the MIT for component number 6, "Review."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	5	50.	**Agreement Coefficient = 1.9**
	2	3	30.	
	3	0	0	
	4	2	20.	
	5	0	0	
	6	0	0	
	Total	10	100%	

7. For component number 7, "Summary."

Results show that six items (60%) had perfect agreement; one item (10%) had all ratings contained in two adjacent response categories; one item (10%) had all ratings contained in three adjacent response categories; two items (20%) had all ratings contained in four adjacent response categories; and there were no items that needed five or six adjacent response categories to contain all of the ratings.

Table 10 - Agreement coefficient among the three members of the MIT for component number 7, "Summary."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	6	60.	**Agreement Coefficient = 1.9**
	2	1	10.	
	3	1	10.	
	4	2	20.	
	5	0	0	
	6	0	0	
	Total	10	100%	

8. For component number 8, "Homework."

Results show that six items (67%) had perfect agreement; two items (22%) had all ratings contained in two adjacent response categories; one item (11%) had all ratings contained in three adjacent response categories; and there were no items that needed four, five or six adjacent response categories to contain all of the ratings.

Table 11 - Agreement coefficient among the three members of the MIT for component number 8, "Homework."

Raters	Agreement Values	Number of Items	Percent of Items	
MIT	1	6	66.66	**Agreement Coefficient = 1.44**
	2	2	22.2	
	3	1	11.1	
	4	0	0	
	5	0	0	
	6	0	0	
	Total	9	100%	

Table 12 is a compilation of the coefficients of agreements from Tables 4-11.

Table 12 - Summary of the coefficients of agreement among MIT members across both continuous and categorical responses on each component.

Component	Agreement Coefficient
1. Equipment/materials	1.75
2. Gym setup	1.75
3. Preparation of students	2.00
4. Explanation/demonstration	1.35
5. Practice	1.46
6. Review	1.90
7. Summary	1.90
8. Homework	1.44

Continuous responses for the components of the instrument

The first two components (equipment/materials and gym setup) were combined because each contained only four items and they are similar in their content. Results of the analysis of continuous data obtained for the components follow.

For components #1 & #2 (equipment/materials and gym setup):

Mean inter-rater reliability among MIT members = .86

For component #3 (preparation of students):

Mean inter-rater reliability among MIT members = .83

For component #4 (explanation/demonstration):

Mean inter-rater reliability among MIT members = .89

For component #5 (practice):

Mean inter-rater reliability among MIT members = .68

For component #6 (review):

Mean inter-rater reliability among MIT members = .62

For component #7 (summary):

Mean inter-rater reliability among MIT members = .45

For component #8 (homework):

Mean inter-rater reliability among MIT members = .83

Step 6: Establish the Content Validity of the Instrument

Overview

The purpose of this step was to establish the content validity of the instrument. Information obtained for this step were obtained from the MIT members. The following two basic questions provided the framework for judging content validity:

1. Do the components of the instrument appropriately represent the components of the MI-EPEC K-2 lessons?
2. Do the items contained in each component of the instrument appropriately represent teacher implementation behaviors needed to properly implement MI-EPEC K-2 lessons?

The MIT members were asked to make judgments relative to the above two questions by reviewing the revised instrument and circling the response (Yes or No) for each component and each item of the instrument that best represented their expert opinions. When their response was "No" they were asked to identify the changes necessary to allow them to respond "Yes" in a subsequent rating. A summary of the results follows.

Results

For the instrument's representation of the components included in the MI-EPEC lessons, all MIT members responded "Yes" to all components. For the items contained in

each component of the instrument, results show that all MIT members responded "Yes" to all items. Their responses included editorial changes for 18 items.

Representative examples of these editorial suggestions were: add the letter "s," delete double space errors, change the verb from present to the past. There was a suggestion to define the term "space" in the following item "The space (per student) used at each station for equipment, materials and students matched what was written in the lesson." The definition term "per student" was added parenthetically after the term space.

Also there were suggestions to reword six items as follows:

1. The original item, "The kind of teaching stations specified in the lesson component (distances, target sizes, available space . . .) were used," was restated to read "The kind of teaching stations specified in the lesson component were used."
2. The original item, "The teachers facility was sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson," was restated to read "The teaching facility was sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson."
3. The original item, "The teacher used the key action verbs (tell, ask, restate, etc.) that were described in the lesson," was restated to read "The teacher applied the key action verbs (tell, ask, restate, etc.) that were described in the lesson."
4. The original item, "The transition occurred in the order described in the lesson," was restated to read "The events making up the transition occurred in the order described in the lesson."
5. The original item "If applicable (when called by the lesson) students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated," was restated to read "When called by the lesson students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated as written."

6. The original item, "The teacher was positioned where all students could participate in the explanation/demonstration," was restated to read "The teacher was positioned where all students could benefit equally from the explanation/demonstration."

Step 7: Establish the Reliability of the Instrument

Overview

The purpose of this step was to establish inter-rater reliability for the instrument when measurements obtained were derived from a sample representative of potential users. This process involved selecting a four student rating team (SRT) from the University of Michigan who were involved in a related assessment project, conducting an inservice education session devoted to the what, why and how to use the instrument and having this group rate a videotape of a teacher implementing a randomly selected K-2 lesson (grade 2, lesson 26).

The SRT was comprised of four students from the University of Michigan enrolled in the Department of Kinesiology. Three were graduate students, one each in the area of exercise physiology, medicine (movement science as undergraduate), and nutrition/physical education. The fourth was a senior, undergraduate student in exercise and sport science. All were experienced in evaluating motor performance of elementary age children via videotape. None, however had experience in teaching or first-hand knowledge of MI-EPEC.

A one-day training session was planned and conducted for the SRT. Dr. Paul Vogel, who volunteered to help the investigator, conducted the inservice training session. The agenda for the training is included in Appendix G. Subsequent to the training session, the SRT members used the instrument to rate the videotape.

Results

Rating responses obtained from the SRT were used to calculate the inter-rater reliability of the instrument. Initial analysis was devoted to items where responses were continuous in

nature (items that received a NA or NR response by one or more raters were analyzed separately as categorical data). Descriptive statistics and SPSS (Pearson correlation coefficient) were used to analyze the data. Results of the analysis are presented for the instrument as a whole, for the components of the instrument, and for the items of the instrument.

Descriptive statistics for the instrument as a whole

To portray the amount of agreement among raters across both continuous and categorical responses for the SRT, the ratings were analyzed by response category. These data are included in Table 13. The table located in Appendix N (Degree of Agreement Among Members of the SRT) shows the degree of agreement across all items (column 1), the continuous responses 1-5 by response category in column 3-7, the NA and NR responses, the criterion score established by the MIT in step 5.2, and agreement on numerical (continuous) responses and agreement across all responses in the last two columns. The data in Appendix N are similar in format to those data reported in step 5.2 for the MIT.

A "coefficient of agreement" (COA) was created to portray the overall agreement in ratings among members of the SRT. The COA is the mean of the agreement values (1, 2, 3, 4 and 5) portrayed in the last two columns of the table located in Appendix N. A summary of the results of these calculations are presented in Table 13.

Table 13 - Summary of the degree of agreement among the four members of the SRT on continuous and categorical data.

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	27	(35)	**Agreement Coefficient = 2.473**
	2	12	(16)	
	3	16	(21)	
	4	16	(21)	
	5	5	(7)	
	6	0	(0)	
	Total	76	100%	

The ratings show that 27 items (35.5%) had perfect agreement; 12 items (15.8%) had all ratings contained in two adjacent response categories; 16 items (21.05%) had all

ratings contained in three adjacent response categories; 16 items (21.05%) had all ratings contained in four adjacent response categories and five items (6.6%) had all ratings contained in five adjacent response categories. There were no items that needed six adjacent response categories to contain all of the ratings.

Continuous responses for the instrument as a whole

Analysis of the rating responses which were continuous in nature (1-5) included a mean inter-rater reliability among the SRT members of .35. As with the inter-rater reliability calculated for the MIT, this correlation coefficient represents the average of the correlations of all possible pairs of raters using SPSS and it's Pearson product moment function. To assess the relationship between the SRT scores and the criterion scores obtained from the MIT for continuous items, the same analysis was run. The result obtained between four members of the SRT and the criterion score was .52.

Categorical responses for the instrument as a whole

The rating responses obtained from the SRT, that were categorical (NA and NR) were analyzed by coding all continuous responses as "1" and all categorical responses as "2," and then running the same SPSS Pearson correlation. Results of this analysis follow.

The mean inter-rater reliability among the SRT members was .89 and the mean correlation between SRT members and the criterion score was = .90

Descriptive responses for the components of the instrument

To portray the amount of agreement among raters across both continuous and categorical responses for each component of the instrument SRT ratings were analyzed by response category within components of the instrument. These data are presented by component in tables 14-21 and summarized in Table 22.

Again, a "coefficient of agreement" (COA) was created to portray the overall agreement in ratings among members of the SRT on each of the eight components.

The degree of agreement across all responses were as follows:

1. For component number 1, "Equipment/materials."

   Results show that three items (75%) had perfect agreement; one item (25%) had all ratings contained in four adjacent response categories. There were no items that needed two, three, five or six adjacent response categories to contain all of the ratings.

Table 14 - Agreement coefficient among the four members of the SRT for component number 1, "Equipment/materials."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	3	75.	**Agreement Coefficient = 1.75**
	2	0	0	
	3	0	0	
	4	1	25.	
	5	0	0	
	6	0	0	
	Total	4	100%	

2. For component number 2, "Gym setup."

   Results show that there were no items that achieved perfect agreement; two items (50%) had all ratings contained in three adjacent response categories; and two items (50%) had all ratings contained in four adjacent response categories. There were no items that needed two, five or six adjacent response categories to contain all of the ratings.

Table 15 - Agreement coefficient among the four members of the SRT for component number 2, "Gym setup."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	0	0	**Agreement Coefficient = 3.5**
	2	0	0	
	3	2	50.	
	4	2	50.	
	5	0	0	
	6	0	0	
	Total	4	100%	

3. For component number 3, "Preparation of students."

Results show that two items (20%) had perfect agreement; two items (20%) had all ratings contained in two adjacent response categories; five items (50%) had all ratings contained in three adjacent response categories; one item (10%) had all ratings contained in four adjacent response categories; and there were no items that needed five or six adjacent response categories to contain all of the ratings.

Table 16 - Agreement coefficient among the four members of the SRT for component number 3, "Preparation of students"

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	2	20.	**Agreement Coefficient = 2.5**
	2	2	20.	
	3	5	50.	
	4	1	10.	
	5	0	0	
	6	0	0	
	Total	10	100%	

4. For component number 4, "Explanation/demonstration."

Results show that one item (7%) had perfect agreement; two items (14%) had all ratings contained in two adjacent response categories; seven items (50%) had all ratings contained in three adjacent response categories; three items (21.4%) had all ratings contained in four adjacent response categories; and one item (7%) had all ratings contained in five adjacent response categories. There were no items that needed six adjacent response categories to contain all of the ratings.

Table 17 - Agreement coefficient among the four members of the SRT for component number 4, "Explanation/demonstration."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	1	7.14	**Agreement Coefficient = 3.07**
	2	2	14.28	
	3	7	50.00	
	4	3	21.4	
	5	1	7.14	
	6	0	0	
	Total	14	100%	

5. For component number 5, "Practice."

Results show that six items (40%) had perfect agreement; three items (20%) had all ratings contained in two adjacent response categories; one item (7%) had all ratings contained in three adjacent response categories; three items (20%) had all ratings contained in four adjacent response categories; and two items (13%) had all ratings contained in five adjacent response categories. There were no items that needed six adjacent response categories to contain all of the ratings.

Table 18 - Agreement coefficient among the four members of the SRT for component number 5, "Practice."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	6	40.	**Agreement Coefficient = 2.46**
	2	3	20.	
	3	1	6.66	
	4	3	20.	
	5	2	13.3	
	6	0	0	
	Total	15	100%	

6. For component number 6, "Review."

Results show that five items (50%) had perfect agreement; one item (10%) had all ratings contained in two adjacent response categories; one item (10%) had all ratings contained in three adjacent response categories; two items (20%) had all ratings contained in four adjacent response categories; and one item (10%) had all ratings contained in five adjacent response categories. There were no items that needed six adjacent response categories to contain all of the ratings.

Table 19 - Agreement coefficient among the four members of the SRT for component number 6, "Review."

Raters	Agreement Values	Number of items	Percent of items	
SRT	1	5	50.	**Agreement Coefficient = 2.3**
	2	1	10.	
	3	1	10.	
	4	2	20.	
	5	1	10.	
	6	0	0	
	Total	10	100%	

7. For component number 7, "Summary."

Results show that four items (40%) had perfect agreement; two items (20%) had all ratings contained in two adjacent response categories; three items (30%) had all ratings contained in four adjacent response categories; and one item (10%) had all ratings contained in five adjacent response categories. There were no items that needed three or six adjacent response categories to contain all of the ratings.

Table 20 - Agreement coefficient among the four members of the SRT for component number 7, "Summary."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	4	40.	**Agreement Coefficient = 2.5**
	2	2	20.	
	3	0	0	
	4	3	30.	
	5	1	10.	
	6	0	0	
	Total	10	100%	

8. For component number 8, "Homework."

Results show that six items (67%) had perfect agreement; two items (22%) had all ratings contained in two adjacent response categories; and one item (11%) had all ratings contained in four adjacent response categories. There were no items that needed three, five or six adjacent response categories to contain all of the ratings.

Table 21 - Agreement coefficient among the four members of the SRT for component number 8, "Homework."

Raters	Agreement Values	Number of Items	Percent of Items	
SRT	1	6	66.66	**Agreement Coefficient = 1.55**
	2	2	22.20	
	3	0	0	
	4	1	11.10	
	5	0	0	
	6	0	0	
	Total	9	100%	

Table 22 is a compilation of the coefficients of agreement from Tables 14-21.

Table 22 - Summary of the coefficients of agreement among SRT members across both continuous and categorical responses on each component.

Component	Agreement Coefficient
1. Equipment/materials	1.75
2. Gym setup	3.50
3. Preparation of students	2.50
4. Explanation/demonstration	3.07
5. Practice	2.46
6. Review	2.30
7. Summary	2.50
8. Homework	1.55

Continuous responses for the components of the instrument

The first two components (equipment/materials and gym setup) were combined because each one of them contained only four items and they are similar in their content. Results of the analysis of continuous data obtained for the components follow.

For components #1 & #2 (Equipment/materials and Gym setup):

Mean correlation between SRT members and the criterion = .34

Mean inter-rater reliability among SRT members (i.e., mean correlation among SRT ratings) = .13

Number of items included in the analysis = 7

For component #3 (Preparation of students):

Mean correlation between SRT members and the criterion = .68

Mean inter-rater reliability among SRT members (i.e., mean correlation among SRT ratings) = .54

Number of items included in the analysis = 9

For component #4 (Explanation/demonstration):

Mean correlation between SRT members and criterion = .51

Mean inter-rater reliability among SRT members (i.e., mean correlation among student ratings) = .30

Number of items included in the analysis = 10

For component #5 (Practice):

Mean correlation between SRT members and the criterion = .65

Mean inter-rater reliability among SRT members (i.e., mean correlation among student ratings) = .44

Number of items included in the analysis = 9

For component #6 (Review):

Mean correlation between SRT members and the criterion = .48

Mean inter-rater reliability among SRT members (i.e., mean correlation among student ratings) = .39

Number of items included in the analysis = 6

For component #7 (Summary):

Mean correlation between SRT members and the criterion = .32

Mean inter-rater reliability among SRT members (i.e., mean correlation among student ratings) = .04

Number of items included in the analysis = 5

For component #8 (Homework):

Mean correlation between SRT members and the criterion = .68

Mean inter-rater reliability among SRT members (i.e., mean correlation among student ratings) = .55

Number of items included in the analysis = 3

Categorical responses for the components of the instrument

No correlations were calculated for the categorical responses for each component because of the small number of items that were rated NA or NR in each component.

CHAPTER V

Discussion of the Results, Conclusions and Recommendation

The purpose of this chapter is to discuss the results presented in Chapter 4, draw conclusions, and suggest recommendations related to the steps used in developing the instrument and/or for a follow up study. Each step (and sub-step, where appropriate) used in the process of developing the instrument is presented in this chapter by using the following format: overview, discussion, conclusions and recommendations. Recommendations for the next draft of the instrument are then summarized at the end of this chapter.

Step 1: Determine the Purpose and Objectives of the Instrument

Overview

The purpose of step 1 was to create a clear statement of purpose and objectives for the instrument. The statement of purpose and objectives was viewed as an important guide to the rest of the development process.

Four sub-steps were used to complete step 1. They were: 1.1) collect information related to the instrument's purpose and objectives from the MIT; 1.2) create a prototype purpose statement and objectives; 1.3) obtain feedback from the MIT on the quality of the prototype statements; and 1.4) finalize the instrument's statement of purpose and objectives. This step is consistent with the literature on instrument development and particularly those devoted to assessment (McDavis, 1976; Horine, 1981; Meng and Doran, 1993).

Overview of sub-step 1.1

The purpose of sub-step 1.1 was to obtain information from the MIT members that would enable draft prototype statements of the purpose and objectives of the instrument to be created. Completion of this sub-step was done as follows:

1. Members of the MIT were asked to answer the five questions included in the instrument contained in Appendix A.
2. Data obtained from each member was summarized by question.
3. Summarized statements were then organized according to their commonalties.

Discussion

Data obtained from the MIT members related to the five questions was similar substance. Because of the similarity of responses, they were categorized based on their commonalties under each question. The 71 statements obtained were categorized into the 24 summary statements as reported in Chapter 4.

As noted in Chapter 4, all of the obtained statements were also categorized for their utility in creating the statement of purpose and/or objectives for the instrument. The categorization of the 71 statements into 24 representative statements was helpful in the process of creating a single purpose statement and a set of discrete objectives to guide the development process. It was also helpful, however, to review the original 71 statements at various points in the creation of the draft purpose and objective statements.

Conclusions

The results obtained in sub-step 1.1 were helpful in the creation of the instrument's purpose and objective statements.

The method used provided the information necessary to create a draft purpose statement and objectives sufficiently appropriate to be refined in subsequent steps. Accordingly, the methodology used appears to be appropriate. The MIT members also reported that the instructional materials design criteria provided with the step 1.1 instrument were helpful in creating their responses to the questions.

Recommendations

Based on the results obtained, the following recommendations are made:

1. The method used to obtain information necessary to create draft purpose and objectives statements should be considered for other similar studies.
2. Subjects in similar studies should be intimately familiar with the subject matter of the study.
3. Questions of the type used in step 1.1 are appropriate for obtaining information leading to the development of prototype purpose and objectives.
4. Although more subjects would likely result in more responses, which in turn could lead to more information to support the development of a statement of purpose and objectives, it is unknown if the number and quality of summary statements obtained would be changed.
5. Providing the members of the MIT with specific criteria to guide the design of the instructional materials was helpful in generating responses to the questions. Accordingly, it is recommended for similar studies.

Overview of sub-step 1.2

The purpose of sub-step 1.2 was to create a prototype statement of purpose and objectives to guide development of the instrument.

Discussion

The 24 summary categories, in combination with the 71 responses to the four questions of step 1.1 (see Appendix B) and their organization by question (see pages 65-67) were very helpful in creating the purpose statement and objectives of the instrument. Review of statements coded with a "P" (Purpose) provided the information necessary to create a purpose statement that contained appropriate content, key words and/or phrases and a "standard" to assure comprehensiveness. With the "substance" of the statement covered, it was easy to address issues of "brevity," and "clarity".

Conclusions

The procedures used to develop the purpose and objectives of the instrument (assessing the possibility of use of each statement obtained) were appropriate and helpful.

They provided the information necessary to establish the appropriate content, which in turn provided a standard for assuring that work on the technical quality of the statements did not erode the quality of their substance.

Recommendations

1. The procedures used to develop a draft statements of purpose and objectives were appropriate; they should be considered for use in similar studies.
2. Organizing the data into similar categories and labeling statements for their potential use in drafting the purpose and/or objectives statements should also be considered for other studies designed to produce draft statements of purpose and objectives.

Overview of sub-step 1.3 and 1.4:

The purpose of sub-steps 1.3 and 1.4 was to review and finalize the statement of purpose and objectives created as a result of sub-steps 1.1 and 1.2. Sub-step 1.3 was completed by asking the MIT members to review the appropriateness of the draft prototype statements and offer suggestions, additions, deletions and/or alterations to the drafts. They were asked to sign off on the revised statement of purpose and objectives in sub-step 1.4.

Discussion

This review process resulted in significant improvements to the statements. They included changes in words, adding adjectives and phrases and deleting and refining entire objectives. The inclusion of a request for rationales for each suggested change were particularly helpful. These statements were helpful in framing the discussion in the meeting with the MIT members.

Conclusions

1. The procedures used to review and revise the draft statements of purpose and objectives were appropriate for this project.
2. Because of the extensive background of members of the MIT and their commitment to the evaluation process in general, it is impossible to state how effective this procedure would be with other, perhaps less committed subjects.

Recommendations

1. The procedures used in this step should be considered for use in similar studies designed to refine directional statements like the purpose and objectives statements for this instrument.
2. Experienced and professional subjects committed to the purpose of the project are probably necessary to obtain quality statements.

Step 2: Construct a Table of Test Specifications

Overview

Step 2 involved two sub-steps. Sub-step 2.1 was devoted to developing test specifications and sub-step 2.2 to creating an item pool. The purpose of creating test specifications in sub-step 2.1 was to provide a framework for item construction. Sub-step 2.2 was devoted to developing an item pool from the purpose statements and indicators obtained in sub-step 2.1.

Discussion

Thirty purpose statements and 148 indicators were obtained from the MIT members for the nine lesson components. Table 1 in Chapter 4 reveals how these were distributed by lesson component. Review of the purpose statements and indicators obtained for each lesson component (see Appendix H), and comparing them with the literature on effective teaching embodied in the design guidelines used to create MI-EPEC K-2 lessons, reveals a strong match. Review of the number of indicators by each lesson component confirms this match in that the more complex and time consuming lesson components (preparation of the students, explanation/demonstration and practice) included a higher number of indicators

The test development literature states that different sources can be used to construct the items of an instrument. Seidman et al. (1979) advocated selection of items based on content considerations. In terms of developing the item pool, Wotruba and Wright

(undated) identified two methods for developing an item pool. One includes searching the relevant literature and the other involves interviewing stakeholders.

This study incorporated both methods in that the design criteria provided to the MIT members in steps 1 and 2 represent the literature on effective teacher behavior and lesson design, and the MIT members represent highly informed stakeholders in the process. Results of sub-step 2.1 were the basis for developing an item pool in sub-step 2.2. Applying the criteria of measurability, discreteness, clarity and lack of redundancy to the indicators resulted in a pool of items to evaluate teacher implementation behaviors suitable for review by the MIT members in step 3 of the instrument development process.

Conclusions

The test specification process used in step 2 provided a suitable framework for item construction. Based on the results obtained in this step, we can conclude that the narrative form used for the creation of an item pool representative of important MI-EPEC teacher implementation behaviors was appropriate for obtaining indicators for each lesson component.

Recommendations

1. The methods used in this step to develop a pool of items should be considered for use in similar studies.
2. Participants in the process should be provided with materials which clearly specify desired implementation behavior to help them create relevant purpose statements and indicators for each lesson component.
3. It is important that participants in this process know the guidelines by which the instructional intents are to be achieved and be committed to the need for the product being produced.

Step 3: Review and Revise the Item Pool

Overview

The purpose of this step was to obtain expert review of the prototype items created in sub-step 2.2 and to obtain suggestions for revisions that would include supporting rationales. Members of the MIT were asked to rate the appropriateness and/or inappropriateness of the items and their response format. They also were asked to provide a rationale for changes proposed to transform inappropriate items into appropriate items.

In this regard, the literature stated that one of the development steps of the instrument is reviewing or screening the item pool. The purpose of this step is to reduce the candidate items in number by eliminating items which are obviously redundant, ambiguous and of low importance from the viewpoint of important stakeholders. (McDavis, 1976; Wotruba & Wright, undated).

Discussion

Results obtained from the MIT, (ratings of the items as "appropriate" or "inappropriate" and their rationales) were treated as quantitative and qualitative data. The results presented in Chapter 4 show that there were 40 items (32 appropriate and 8 inappropriate) out of the 115 items included in the item pool on which there was perfect agreement by members of the MIT.

In addition to the 32 items rated "appropriate" by all members of the MIT, there were 38 items rated appropriate by two members, and 36 items rated as appropriate by one member. Accordingly, at least one member of the MIT members rated 106 of the 114 items (93%) as appropriate in this step. The variance in their ratings resulted from a variety of sources. Some were attributed to the differences between rating effective teaching and teaching behaviors written into the MI-EPEC lessons. Others were attributable to the way an item was written and sometimes key words or phrases needed to be clarified or the emphasis altered.

Some variance could be attributed to the anticipated inability of raters to reliably assess a complex teaching behavior. Subsequent to fully discussing the proposed changes (and their associated rationales), all disagreements were resolved.

Conclusions

The procedures used to accomplish step 3 were appropriate. The form designed to collect data from the MIT members regarding the appropriateness of the items provided information to significantly improve the prototype items. Of particular help were the suggested improvements and their statements of rationale. This information quickly focused the discussion during the three meetings and led to item revisions that gained the support of all members of the MIT.

Recommendations

Completion by experts of individual ratings of the appropriateness of items with a request for changes necessary to convert inappropriate items to appropriate items was very helpful. By discussing these suggestions in face-to-face meetings where the suggested improvements and their rationales are explained, appropriate revisions in the items can be made.

Step 4: Prepare Instructions for Using the Instrument

Overview

The purpose of this step was to develop instructions for using the prototype instrument. Draft instructions were developed which included the following parts: introduction, purpose of the instrument, overview of lessons components, response format, and procedures for conducting the measurement process. The initial draft and directions for its review are included in Appendix I. The MIT members were asked to judge the appropriateness of the instructions and suggest how they could be improved.

Discussion

This step was accomplished by providing members of the MIT with draft instructions for use of the instrument. They independently reviewed the appropriateness of the instructions and provided feedback directly on the written instructions.

Results obtained from the MIT members indicated that several changes were necessary to provide users with appropriate information. The changes included suggestions that improved the clarity of the introduction and purpose portions of the instrument. Several changes to the overview of the lesson components were also suggested. It was judged important to clearly describe the nine components of MI-EPEC lessons so that users understand that the instrument is specifically designed to measure the degree to which teachers implement each component.

Revising the order of the description of the response options to begin with "full implementation" and end with "not implemented" was suggested to reinforce the intent of the instrument to measure implementation of MI-EPEC lessons. Accordingly, it was agreed that the initial response should focus on full implementation. Other significant changes in the definition of the scale levels were suggested and supported by the MIT. Rationales for the changes were carefully considered with the intent being to maximize the reliability of the instrument. This dialog also resulted in the addition of a "Not Applicable" (NA) response option for the reason described in Chapter 4. The changes suggested for the measurement procedures section, in addition to those mentioned, are evidence of the need for review by experts of an initial draft of instructions for the instrument.

Conclusions

The procedures used to review and revise the initial draft of the instructions were appropriate. This conclusion is based on the acceptance of the revised draft by the MIT and the significant changes made in the initial draft.

Recommendations

1. Other instrument development projects should at least consider expert review of the instructions for using an instrument.
2. In selecting experts, be sure the group contains strong knowledge of the specific programs instructional materials.

Step 5: Use, Refine and Establish Criterion Scores for the Items of the Instrument

5.1: Use and Refine the Instrument

Overview

This step was designed to pilot test the second draft of the instrument (see Appendix K). The pilot test involved each MIT member using the instrument to rate a videotape of the implementation behavior of a teacher teaching a 15 minute segment of a MI-EPEC lesson. The intent of this step was to:

1. Determine if the items in all components could be used by experts in an authentic assessment setting.
2. Identify difficulties that caused variance in the ratings of individual MIT members.
3. Create a criterion score for each item that could be used during the inservice training of SRT members.

Discussion

It was anticipated that actual use of the instrument in an authentic setting would result in different ratings by members of the MIT. Although the MIT members were instrumental in all phases of instrument development, they had not yet been confronted with rating teacher behaviors that approximate intended implementation behaviors. Nor had they been confronted with alternative behaviors substituted by teachers that fell short or went beyond implementation expectations. Additionally, they did not have the benefit of

practice in rating example teacher implementation behaviors. The issues that emerged resulted in suggestions for changes in the instrument's instructions, rating scale and organization that would resolve the issues identified. Early in the rating process it became appropriate that a "Not Ratable" (NR) response option needed to be added to the rating scale. This response option would allow raters to appropriately score cases where the videotape record was insufficient to judge the degree of implementation of a behavior.

Additional changes, approved for the instructions, rating scale and organization of the transition items, all point to the need for authentic pilot tests as a vital part of instrument development.

Conclusion

At least one pilot test in an authentic situation is critical to the development of a viable teacher implementation behavior instrument suitable for measuring implementation fidelity.

Recommendations

1. Use as many authentic pilot tests as possible during the instrument development process. The pilot tests should be conducted on samples of teacher implementation behavior that is representative of all teacher behaviors expected in actual implementation studies.
2. Early pilot tests should involve evaluators who are very knowledgeable about the instructional intervention, measurement, and evaluation.

Step 5.2: Use, Refine the Instrument, and Establish Criterion Scores for the Items

Overview

After the instrument was revised in step 5.1, it was used a second time by the MIT. The purpose of step 5.2 was to rate teacher implementation behavior on another 15 minute MI-EPEC lesson segment. The intent of this step was to:

1. Identify problems associated with applying the instrument to another teacher and different lesson content.
2. Resolve the problems encountered.
3. Create a criterion score for each item so that the accuracy of the ratings of other users could be determined by comparing the degree to which their scores matched the criterion scores.

Discussion

Review of the data located in Table 3 and Appendix M shows high agreement among the MIT members for the second pilot test of the instrument. All members scored the videotape the same for 44 (58%) of the instrument's items, and for another 20 items (26%), ratings were contained within two adjacent rating categories. The cumulative 84% agreement for MIT ratings falling in one or two adjacent rating categories is perceived as very high for rating complex teaching behavior. The coefficient of agreement (COA = 1.66) calculated from the data included in the table located in Appendix M reflects close agreement of the MIT members in scoring this videotape. This coefficient represents the fact that, on average, the MIT ratings fell between perfect agreement and having all their scores contained within two adjacent response categories across all 76 items of the instrument. Even though the data show high agreement between the MIT members additional changes were agreed upon by the MIT that would improve agreement in subsequent assessments.

It is not possible to determine the extent of future changes that should be made to the instrument from the data obtained in this study. Each replication of the assessment process could provide rating difficulties not encountered in prior assessments. Using the instrument to rate additional teachers implementing other lessons could answer this question but this is beyond the scope of this study. It should also be noted that whole classes of item changes may be necessary to extend the instrument's use to the cognitive, affective and fitness content, which is also part of MI-EPEC lessons.

Conclusions

1. Use of the instrument a second time was important for refining the instrument.
2. It was possible to establish criterion scores for each item and the process of establishing these scores led to additional changes in the items of the instrument.
3. Agreement among the MIT members using the revised instrument to rate a second videotape was high.
4. Provided raters are knowledgeable of the purpose, intent, structure and design criteria for MI-EPEC K-2 lessons and are aware of the need, purpose and characteristics of evaluating teacher implementation behavior, reliable ratings of teacher implementation behavior can be obtained for teachers similar to those selected for pilot testing on MI-EPEC K-2 lessons.

Recommendations

The results obtained in this study suggest that development of this instrument should continue if documentation of teacher implementation behavior in implementing MI-EPEC K-2 lessons is needed as part of a broad evaluation plan for the project. It would be appropriate to evaluate additional teachers on additional motor skills, at various places in the teaching-learning progressions of MI-EPEC's structure. It is also important to extend development to content in the other goal areas of the program.

Step 6: Establish the Content Validity of the Instrument

Overview

The purpose of this step was to establish the content validity of the instrument. Data for this step were obtained from the MIT members. Two basic questions provided the framework for judging content validity:

1. Do the components of the instrument appropriately represent the components of the MI-EPEC K-2 lessons?

2. Do the items contained in each component of the instrument appropriately represent teacher implementation behaviors needed to properly implement MI-EPEC K-2 lessons?

This step was accomplished by using the form included in Appendix F. The MIT members were asked to judge each component and each item included in the instrument by responding "Yes" or "No" and stating their rationale in the case of "No" responses.

Discussion

The results obtained from the MIT ratings of the appropriateness of the instrument's components when compared to the lesson's components show that the instrument's components are well matched to the components of the MI-EPEC lessons. This can be attributed to incorporating the MI-EPEC design criteria into the selection of the instrument's components from the beginning of the instrument development process.

There was also agreement among all MIT members on all of the items included in this draft of the instrument. Although there were editorial suggestions for six items, the content of the items was determined to be valid by this group. The results obtained can, in part, be attributed to the use of MI-EPEC design criteria in the development. It can also be argued that the intensive use of the MIT in the development of the instrument, including estimating content validity, would naturally result in high estimates of content validity due to the vested interest. Although this limitation cannot be refuted at this time, subsequent content validity studies can be conducted using others who are expert in MI-EPEC instructional materials, the need for evaluation data in the curriculum development process and the instrument development process.

Conclusions

According to the MIT, the content of the instrument is valid. It represents MI-EPEC lesson components and the items within the components measure teacher behaviors incorporated into the MI-EPEC K-2 lessons. Accordingly, it can be concluded that the procedures used to create the instrument were appropriate.

Recommendations

Because of the extensive involvement of the MIT members in steps 1-6 of the development process, it would be appropriate to replicate the content validity step with another group of content experts.

Step 7: Establish the Reliability of the Instrument

Overview

The purpose of this step was to establish inter-rater reliability for the instrument when measurements obtained were derived from a sample representative of potential users.

A one-day training session was conducted for the SRT. The agenda for the training is included in Appendix G. Subsequent to the training session, SRT members used the instrument to rate a videotape of a teacher implementing a K-2 lesson segment devoted to catching balls.

Discussion

The results obtained from analyzing the rating responses of the SRT provide an indication of the reliability of the instrument when used by relatively naive raters. The procedure used was consistent with those used by Smith, Smoll and Hunt (1977), Weston, Petosa and Pate (1997), Stallings (1978) and Ruud (1976), who also used experts and other observers to independently evaluate the behaviors of interest. The authors consulted each other to create criterion scores which in turn were used to assess the accuracy of the other observers. Reliability coefficients were computed between all possible pairs of observers using the Pearson correlation coefficient.

Reliability of the instrument as a whole

Results contained in Table 13 show that on average 2.5 adjacent response categories were necessary to contain the ratings of the SRT. When compared to the data in Table 3 for the MIT (COA = 1.7), the variance is much greater. All of the data indicated that individuals with extensive background in the development of the MI-EPEC

instructional materials and who were involved in the development of the implementation instrument produced more reliable ratings than individuals without such background.

The results obtained from the correlational analysis of the continuous items support the same conclusion. For these data, the inter-rater reliability for the MIT and SRT was .77 and .35 respectively. For items analyzed using categorical data the MIT correlation was .96 and the SRT correlation .89. When the SRT responses were compared with the criterion scores the correlation was .52.

There is no way to determine from the data whether the differences in reliability between the MIT and SRT could be reduced by using selection criteria favoring more experienced raters and/or by improving the effectiveness of the inservice training. However, the four correlation coefficients between the SRT members and the criterion scores (.708, .415, .405 and .560) show that one member of the SRT had a relatively high correlation coefficient (.708). This suggests selection may be important to obtaining reliable ratings. This member of the SRT was the only one who had a major in movement science (biomechanics).

It should be noted here that the one-day inservice did not provide enough time for the SRT members to practice and demonstrate mastery in using all items contained in the instrument. In fact, it was clear that more rating practice was needed. Due to the pre established constraints which guide the training, this was not possible to correct.

Reliability related to the components of the instrument

The coefficients of agreement obtained for each component of the instrument among continuous and categorical responses of the MIT reveal that eight components were equal to or below 2.0. The coefficients of agreement provide an indication of strong, consistent agreement. This was not true, however, for the SRT, where the coefficients of agreements ranged from 1.55 (homework) to 3.5 (gym setup).

Conclusions

1. Individuals who are very familiar with MI-EPEC and who were involved in developing the teacher implementation behavior instrument can reliably assess the degree to which teachers implement MI-EPEC K-2 lessons devoted to instruction on motor skills.
2. Individuals who are less familiar with MI-EPEC and who were not involved in the development of the instrument are not able to reliably rate teacher implementation behavior.
3. It is important to conduct rater education that is sufficient to provide demonstratable competence in matching expert-established criterion scores prior to scoring videotapes of teacher implementation behavior.

Recommendations

1. Review of the results obtained and observation of the inservice training of the SRT members suggest that more time is needed to practice using the instrument during the inservice training. It is important for raters to demonstrate their ability to match criterion scores prior to rating videotapes of teacher implementation behavior.
2. It is also important to identify characteristics of raters that relate to developing rating competence.
3. The early trials and the reliability scores obtained for the MIT suggest that this instrument shows promise for assessing teacher implementation behavior of MI-EPEC lessons. Work on the instrument should be continued and expanded to the other content areas of MI-EPEC.

Discussion of the Study

The procedural steps used to develop the Teacher Implementation Evaluation Instrument resulted in an instrument that has considerable promise for measuring key aspects of implementation behavior of teachers teaching MI-EPEC K-2 lessons. The strong connection of the instrument's developmental steps to MI-EPEC criteria used to

design K-2 lessons no doubt limits the use of the instrument for assessing implementation of other instructional materials.

The instrument would not be suitable for general physical education use without substantial changes designed to match items with specific intent with the instructional materials of interest. Accordingly, there would probably be changes needed even if the instrument were to be used for MI-EPEC 3-5 materials.

Most other instructional materials available to teachers or researchers are not sufficiently detailed to allow for replication across teachers and implementation contexts. Where replication of effective instruction is not possible, there is little purpose in measuring teacher implementation behavior or, for that matter, in evaluating effectiveness of instructional materials to affect student achievement.

The instrument as it is presently stands needs additional improvement. In addition to the recommendations stated above, there are other recommendations listed below for the next draft of the instrument.

1. Refine the process of selecting and training potential users of the instrument.
    1.1 Train raters to mastery. At the completion of the training, raters must be able to demonstrate competence before being permitted to collect data.
    1.2 Seek to understand what variables are associated with raters obtaining rating competency and develop rater selection criteria.
2. Expand the subjects and number of lessons used to calculate reliability and validity for the instrument as a whole and for its component parts.
3. Continue work on reducing the variability of ratings on individual items of the instrument.
4. Fully develop, test and refine the procedures for documenting the degree to which teachers implement lesson components in accordance with the instructional time intervals specified in the lesson materials.
5. Use item analysis procedures to attempt to reduce the size of the instrument.

6. Create a scoring system that accounts for the relative importance of the instruments items.
7. Expand the use of the instrument to provide for measuring the implementation behaviors of teachers teaching MI-EPEC K-2 lessons to include the lesson materials devoted to fitness, personal/social/attitudinal and cognitive objectives.

REFERENCES

Barlow, D. H., & Hersen, M. (1984). Single case experimental designs: Strategies for studying behavior change (2nd ed.). Elmsford, NY: Pergamon Press.

Blair, S. N., Koh, H. W., Paffenbarger, R. S., Clark, D. G., Cooper, K. H., & Gibbons, L. W. (1989). Physial fitness and all-cause mortality. A prospective study of healthy men and women. JAMA, 262(17), 2395-2401.

Broome, S. A., & White, R. B. (1995). The many uses of videotape in classrooms serving youth with behavioral disorders. Teaching Exceptional Children, 27, 10-13.

Brophy, J. (1980). Recent research on teaching (Occasional Paper No. 40). East Lansing, MI: Michigan State University, The Institute for Research on Teaching.

Carpenter, J. O. (1977). Competence—Effectiveness of instruction in higher education and how to objectively assess it and improve it. Lynchburg, VA: Center for the Improvement of Instruction and Learning.

Center for Disease Control. (1987). Protective effect of physical activity on coronary heart disease. MMWR, 36.

Colbert, J. O. (1977). The development and implementation of the Levels of Use Observational Inventory (LoUOI): An instrument to aid in the adoption of an innovation process. Paper presented at the annual meeting of the national association for research in science teaching, Cincinnati, OH.

Crews, D. J., & Landers, D. M. (1987). A meta analytic review of aerobic fitness and reactivity to psychosocial stressors. Medi Science Sports Exercise, 19, s114-s120.

Darst, P. W., Mancini, V. H., & Zakrajsek, D. B. (1983). Systematic observation instrumentation for physical education. Champaign, IL: Leisure Press.

Darst, P. W., Zakrajsek, D. B, & Mancini, V. H. (1980). Analyzing physical education and sport instruction. Champaign, IL: Leisure Press.

DiNucci, J. (1988). Measurement for evaluation by physical education professionals. Edina, MN: Burgess International Group, Inc., Bellwether Press Division.

Dummer, G. M., Reuschlein, P. L., Haubenstricker, J. H., Vogel, P. G., & Cavanaugh, P. L. (1993). Evaluation of K-12 physical education programs: A self study approach. Dubuque, IA: Brown and Benchmark.

Educational and Cultural Center. (1969). Behavioral analysis instrument for teachers. Syracuse, NY: US Department of Health, Education and Welfare, Office of Education.

Ennis, C. D. (1992). Developing a physical education curriculum based on learning goals. Journal of Physical Education, Recreation and Dance, 63(7).

Fargo, S. J., Taylor, J. R., & Peterson, D. L. (1990). Arizona Lesson Observation and Evaluation (ALOE). Flagstaff, AZ: Northern Arizona University, Division for Research and Faculty Development, Center for Excellence in Education.

Fletcher, J. L., & Spady, W. G. (1975). The development of instrumentation to measure the alternative operational manifestations of five basic functions of schooling. Paper prepared for delivery at the American Educational Research Association Annual Meeting, Washington, DC.

Furst, N., & Hill, R. (1971). Classroom observation interaction. In L. Deighton (Ed.), Encyclopedia of education. New York: Macmillan.

Good, C. (Ed.). (1959). Dictionary of education, (2nd ed.). New York: McGraw-Hill.

Gunter, P. L., Jack, S. L., Shores, R. E., Carrel, D., & Flowers, J. (1993). Lag sequential analysis as a tool for functional analysis of student disruptive behavior in classroom. Journal of Emotional and Behavioral Disorders, 1, 138-148.

Harris, S. S., Caspersen, C. J., DeFriese, G. H., & Estes, E. H. (1989). Physical activity counseling for healthy adults as a primary preventive intervention in the clinical setting. The Journal of American Medical Association, 261(24).

Hawkins, A., Wiegand, R. L., & Landin, D. K. (1985). Cataloguing the collective wisdom of teacher educators. Journal of Teaching in Physical Education, 4(4).

Hawkins, R. P., & Dobes, R. W. (1977). Behavioral definitions in applied behavior analysis: Explicit or implicit. In B. C. Etzel, J. M. LeBanc, & D. M. Baer (Eds.), New directions in behavioral research: Theory, methods, and applications. Hillsdale, NJ: Lawrence Erlbaum.

Holland, B. V. (1986). Development and validation of an elementary motor performance test for students classified as non-handicapped, learning disabled or educable mentally impaired. A dissertation submitted to Michigan State University, East Lansing.

Horine, L. (1981). Faculty performance evaluation: One answer to accountability demands. Journal of Physical Education, Recreation and Dance, 52(7).

Jochuas, B., Neubert, N., & Rockman, S. (1979). Assessing the content of instructional television programs for use in formative evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research . In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, IL: Research Press.

Johnston, J. M., & Pennypacker, H. S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Lawrence Erlbaum.

Jordan, D. (1993). Promoting active lifestyles—A multidisciplinary approach. The Journal of Physical Education, Recreation and Dance, 64(1).

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. Journal of Applied Behavior analysis, 10(1).

King, J. A., Morris, L. L., & Fitz-Gibbon, C. T. (1987). How to assess program implementation. Beverly Hills: Sage Publications.

Lakka, T. A. (1994). Relation of leisure-time physical activity and cardiorespiratory fitness to the risk of acute myocardial infarction in men. N Engl J Med., 330, 1549-1554.

Lambert, L. (1987). Secondary school physical education problems: What can we do about them? Journal of Physical Education, Recreation and Dance, 58(2).

Lee, I. M., Hsieh, C. C., & Paffenbarger, R. S. (1995). Exercise intensity and longevity in men: The Harvard Alumni Health Study. The Journal of the American Medical Association, 273(15).

Locke, L. F. (1979). Supervision, schools and student teaching: Why things stay the same. In G. Scott (Ed.) The academy papers (pp. 65-74). Washington DC: The American Academy of Physical Education.

Lopatka, R. (1978). Development of models for assessing affective objectives in the social studies. A report to the Minister's Advisory Committee on student achievement. Alberta Department of Education.

McDavis, R. J. (1976). The development and field testing of an instrument to evaluate student personnel programs. Journal of College Student Personnel, 17(2).

McNeil, J. D., & Popham, W. (1973). The assessment of teacher competence. In. N. Gage (Ed), Second handbook of research on teaching. Chicago: Rand McNally.

Meng, E., & Doran, R. L. (1993). Improving instruction and learning through evaluation: Elementary school science. (ERIC Clearinghouse for Science, Mathematics, and Environmental Education No. ED 359 066)

Morra, F. (undated). Four strategies for collecting information. Unpublished monograph by Frank Morra, Jr.

Morris J. N. (1990). Exercise in leisure time: Coronary attack and death rates. Br Heart J, 63, 325-334.

Morris, J. N., Everitt, M. G., Pollard, R., Chave, S. P., & Semmence, A. M. (1980). Vigorous exercise in leisure time: Protection against coronary heart disease. Lancet, 2(8206), 1207-1210.

Nachmias, D., & Nachmias, C. (1976). Research methods in the social science. New York: St. Martin's Press.

Nieman, D. C. (1989). Exercise: How much is enough? How much is too much?. Journal of Women's Sports and Fitness, 11(5).

Philip, L. G., & Thomas, M. R. (1996). Self evaluation of instruction: A protocol for functional assessment of teaching behavior. Journal of Intervention in School And Clinic, 31(4).

Remmers, H. (1963). Rating methods in research on teaching. In N. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Rink, J. (1979). Development of an observation system for content development in physical education. Unpublished doctoral dissertation, The Ohio State University, Columbus.

Rose, S., Wang, M. C., Maxwell, J., & Corey, E. (1973). The development of a measure to evaluate language communication skills of young children. University of Pittsburgh, Learning Research and Development Center. (ERIC Document Production Service No. ED 076665).

Ruud, O. G. (1976). The construction of an instrument to measure proportional reasoning ability of junior high pupils. A thesis. The University of Minnesota.

Seefeldt, V. (1990). MEAP Test Development Process. Unpublished paper. Kinesiology Department, Michigan State University, East Lansing, MI.

Seefeldt, V. (Ed.) (1986). Physical activity & well-being. Reston, VA: American Alliance for Health, Physical Education, Recreation, and Dance.

Seidman, E., Linney, J. A., Rappaport, J., Herzberger, S., Kramer, J., & Alden, L. (1979). Assessment of classroom behavior: A multiattribute, multisource approach to instrument development and validation. Journal of Educational Psychology, 71(4).

Shearron, G. (1976). Developing and improving instruments for measuring competence. Journal of Teacher Education, (29) 2.

Siedentop, D. (1983). Developing teaching skills in physical education, (2nd ed.) Palo Alto, CA: Mayfield.

Simon, A., & Boyer, E. (1974). Mirrors for behavior III. An anthology of observation instruments. Philadelphia: Communications Materials Center, Research for Better Schools.

Smith, R., Smoll, F., & Hunt, E. (1977). A system for the behavioral assessment of athletic coaches. Research Quarterly, 48, 401-407.

Stallings, J. A. (1978). The development of the contextual observation system. Paper presented to the American Educational Research Association, Toronto, Canada.

Tenebaum, G., Singer, R., & Dishman, R. (1992). Physical activity and psychological benefits. The Physician and Sports Medicine, 20(10).

Voelker, A. M., & Horvat, R. E. (1974). The development of an instrument for determining the nature of elementary school children's environmental decisions. Washington DC: US Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development.

Vogel, P. (1986). Effects of physical education programs on children. In V. Seefeldt (Ed.), Physical activity & well-being (pp. 455-501). Reston, VA: American Alliance for Health, Physical Education, Recreatin, and Dance.

Vogel, P. (1998). Michigan's exemplary physical education curriculum (MI-EPEC). East Lansing, MI: Michigan Fitness Foundation.

Vogel, P., & Seefeldt, V. (1988). Program design in physical education. Indianapolis, IN: Benchmark Press. Inc.

Watson, E. R., Sherrill, A., & Weigand, B. (1994). Curriculum development in a world wide school system. The Journal of Physical Education, Recreation and Dance, 65(1).

Weston, A. T., Petosa, R., & Pate, R. R. (1997). Validation of an instrument for measurement of physical activity in youth. Medicine and Science in Sports and Exercise, 29(1).

Wiemann, J. M. (1981). Effects of laboratory videotaping procedures on selected conversation behaviors. Journal of Human Communication Research, 7(4).

Wittrock, M. C. (1986). Handbook of research on teaching: A project of the American Educational Research Association (3rd ed.). New York: Macmillan Publishing Company.

Wolf, A. E., & Greenwald, M. J. (1978). Developing an observation schedule for coding reading-related activities in the secondary school classroom. A paper presented at the Annual Meeting of the National Reading Conference, St. Petersburg, FL.

Wotruba, T. R., & Wright, P. L. (undated). Developing a teaching effectiveness assessment instrument. San Diego, CA: San Diego State University.

Appendix A

Step 1.1: Implementation Survey (Purpose and Objectives of the Instrument)

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.
C&E: Curriculum and Evaluation Specialists.
E: Evaluation Specialists.

Date:_____

Introduction:

You were a member of the MI-EPEC project team that developed and evaluated physical education lessons for grades K, 1 and 2. To develop an instrument to measure the degree to which teachers use the lessons in accordance with their design, it is important to clearly articulate its purpose and objectives. The purpose and objectives of such an instrument can best be described by individuals involved in the development of the instructional materials. Accordingly, it will be helpful to obtain your responses to the following four questions.

1. Why is it important to measure implementation fidelity?

2. What are you be willing to accept as evidence of high implementation fidelity?

3. What kinds of teacher behaviors would be examples of low implementation fidelity?

4. What variables do you believe will be associated with high and low implementation fidelity?

Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity.

Thank you for your cooperation

Investigator:
Hasan Al-Tawil

Appendix B

Results of Step 1.1

A. Why is it important to measure implementation fidelity?

1. Implementation fidelity is essential to assessing and systematically improving the effects of instruction on students.
2. Maximize the opportunity of students to learn.
3. Maximize positive outcomes on students.
4. Improve the lesson's efficiency.
5. Helps to discover problems with teaching methods, complexity of the task, and/or the willingness of the teacher to implement the lesson as it was written. **(2)**
6. Provides formative information, targeting potential problems and serving as a parameter for potential improvements. **(2)**
7. Allows evaluators to compare potential refinements to intended student outcomes.
8. It is the dependent measure for assessing the effectiveness of inservice training.
9. Helps measure the degree to which teacher implementation behaviors relate to student achievement of intended outcomes.
10. Helps refine inservice training and/or instruction treatments related to student achievement.
11. Defines actual instructional treatment.
12. Assess the effectiveness of instruction by identifying the instruction.
13. Measuring quantitative and qualitative values of the instruction.
14. Identify the degree to which the instruction taught as written.

B. What are you willing to accept as evidence of high implementation fidelity?

1. Teacher's behavior reflects what is written in the lesson. **(2)**
2. Key points are communicated accurately to the students as written in the lesson.
3. Student's behaviors reflect what is written in the lesson.
4. Instructional setting prepared as written in the lesson.
5. Equipment and materials are used as written in the lesson.
6. Teacher implemented the components of the lesson as it is written. **(2)**

7. Teachers used available time as it is allocated in the lesson components. **(2)**
8. Teacher's ability to correctly interpret implementation tasks outlined in the lessons.
9. Each component should be taught accurately with only minor deviations from prescribed set-up, activities and equipment-which do not impact on the instructional intent of the lesson - being tolerated.

C. What kinds of teacher behaviors would be examples of low implementation fidelity?

1. Failing to prepare the activity area as it is described in the lesson. **(2)**
2. Failing to implement components included in the lesson.
3. Failing to communicate key points written in the lesson. **(2)**
4. Engaging the students in activity not written into the lesson. **(2)**
5. Engaging the students in activity not described in the lesson that distracts the students from the lesson objectives.
6. Not directing the instruction toward the objectives of the lesson.
7. Engaging students in objective related activity but either ignoring the specific step(s) of the progression, areas of emphasis noted in the lessons, using key words or their equivalents.
8. Poorly implementing the suggested Teaching Learning Activities (TLAs).
9. Maintaining insufficient discipline to actually teach the lesson.
10. Facing significant interruptions that preclude teaching the lesson.
11. Violating the amounts of time for each lesson component. **(2)**
12. Not implementing all components of the lesson. **(2)**
13. Focusing on a portion of the students versus the whole class.
14. Insufficient practice.
15. Poor feedback (quantity and quality).
16. Insufficient facilities, equipment.
17. Poor communication of lesson content/processes to the students.
18. Insufficient answers given to student questions.
19. Poor/incorrect demonstrations.
20. Disorganization in the presentation of the lesson.
21. Failure to adhere to the essential aspects of appropriate gym setup, equipment, materials, etc.
22. Alterations in the instruction that affect (either positively or negatively) the instructional intent of the lesson.

23. Teaching objectives that are not part of the written lesson.
24. Inability to effectively manage the class environment.

D. What variables do you believe will be associated with high and low implementation fidelity?

1. The way lesson is written (clarity, degree of thoroughness and detail, amount of content).
2. How teachers read for comprehension and their familiarity with the content being addressed. **(2)**
3. The amount of content in the lesson.
4. The teacher's commitment to convey the lesson's learning objective to the students.
5. The degree to which the teacher prepared to implement the lesson.
6. The degree to which the teacher is comfortable with the methods, examples and narrative information used to communicate the intended lesson content.
7. Number of students in the class. **(3)**
8. Nature of the learning environment and expectations of the students prior to implementation of the lesson.
9. Philosophical orientation of the teacher regarding direct instruction. **(2)**
10. Nature (socio-economic status) of the majority of students.
11. Educational expectations the teacher holds for the students.
12. Teacher effectiveness in managing a good learning environment.
13. Teacher motivation levels.
14. Teacher's "comfort" or self-efficacy as it relates to teaching the materials.
15. Incentives for providing "excellent" instruction (from administration, social context, etc.).
16. Teacher preparation time.
17. Number of classes taught (other responsibilities).
18. Amount of contact with class (number of class sessions per week).
19. Availability of sufficient facilities and equipment.
20. Student characteristics (special needs students).
21. Teacher experience level.
22. Single-grade versus multi-grade classes.
23. Support from other faculty/administrators (P.E. and non-P.E. faculty).

E. Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity.

It is important to keep in mind that the terms "implementation fidelity" and "instructional competence" are not synonymous. In fact, there are times that the "best" teachers, with the best ideas, are lowest in implementation fidelity.

Appendix C

Analysis of the Statements for Their Potential Contribution to Creating a Statement of Purpose and/or Objectives for the Instrument

P = Purpose

O = Objective

A.	**Why is it important to measure implementation fidelity?**	P	O
1.	Implementation fidelity is essential to assessing and systematically improving, the effects of instruction on students.	X	X
2.	Maximize the opportunity of students to learn.	X	X
3.	Maximize positive outcomes on students.	X	X
4.	Improve the lesson's efficiency.	X	X
5.	Helps to discover problems with teaching methods, complexity of the task, and/or the willingness of the teacher to implement the lesson as it was written. **(2)**	X	X
6.	Provides formative information, targeting potential problems and serving as a parameter for potential improvements. **(2)**	X	X
7.	Allows evaluators to compare potential refinements to intended student outcomes.	X	X
8.	It is the dependent measure for assessing the effectiveness of inservice training.		X
9.	Helps measure the degree to which teacher implementation behaviors relate to student achievement of intended outcomes.	X	X
10.	Helps refine inservice training and/or instruction treatments related to student achievement.	X	X
11.	Defines actual instructional treatment.		X
12.	Assess the effectiveness of instruction by identifying the instruction.	X	X
13.	Measuring quantitative and qualitative values of the instruction.	X	X
14.	Identify the degree to which the instruction taught as written.	X	X
B.	**What are you willing to accept as evidence of high implementation fidelity?**		
1.	Teacher's behavior reflects what is written in the lesson. **(2)**	X	X
2.	Key points are communicated accurately to the students as written in the lesson.	X	X

	(B statements continued)	**P**	**O**
3.	Student's behaviors reflect what is written in the lesson.	X	X
4.	Instructional setting prepared as written in the lesson.	X	X
5.	Equipment and materials are used as written in the lesson.	X	X
6.	Teacher implemented the components of the lesson as it is written. (2)	X	X
7.	Teachers used available time as it is allocated in the lesson components. (2)	X	X
8.	Teacher's ability to correctly interpret implementation tasks outlined in the lessons.	X	X
9.	Each component should be taught accurately with only minor deviations from prescribed set-up, activities, and equipment -- which do not impact on the instructional intent of the lesson -- being tolerated.	X	X
C.	**What kinds of teacher behaviors would be examples of low implementation fidelity?**		
1.	Failing to prepare the activity area as it is described in the lesson. (2)	X	X
2.	Failing to implement components included in the lesson.	X	X
3.	Failing to communicate key points written in the lesson. (2)	X	X
4.	Engaging the students in activity not written into the lesson. (2)	X	X
5.	Engaging the students in activity not described in the lesson that distracts the students from the lesson objectives.	X	X
6.	Not directing the instruction toward the objectives of the lesson.	X	X
7.	Engaging students in objective related activity but either ignoring the specific step(s) of the progression, areas of emphasis noted in the lessons, using key words or their equivalents.	X	X
8.	Poorly implementing the suggested Teaching Learning Activities (TLAs).	X	X
9.	Maintaining insufficient discipline to actually teach the lesson.		
10.	Facing significant interruptions that preclude teaching the lesson.		
11.	Violating the amounts of time for each lesson component. (2)	X	X
12.	Not implementing all components of the lesson. (2)	X	X
13.	Focusing on a portion of the students versus the whole class.		X
14.	Insufficient practice.	X	X
15.	Poor feedback (quantity and quality).		
16.	Insufficient facilities, equipment.	X	X

	(C statements continued)	P	O
17.	Poor communication of lesson content/processes to the students.		X
18.	Insufficient answers given to student questions.		X
19.	Poor/incorrect demonstrations.	X	X
20.	Disorganization in the presentation of the lesson.		
21.	Failure to adhere to the essential aspects of appropriate gym setup, equipment, materials, etc.	X	X
22.	Alterations in the instruction that affect (either positively or negatively) the instructional intent of the lesson.	X	X
23.	Teaching objectives that are not part of the written lesson.	X	X
24.	Inability to effectively manage the class environment.		
D.	**What variables do you believe will be associated with high and low implementation fidelity?**		
1.	The way lesson is written (clarity, degree of thoroughness and detail, amount of content).	X	
2.	How teachers read for comprehension and their familiarity with the content being addressed. **(2)**		
3.	The amount of content in the lesson.	X	X
4.	The teacher's commitment to convey the lesson's learning objective to the students.		
5.	The degree to which the teacher prepared to implement the lesson.		
6.	The degree to which the teacher is comfortable with the methods, examples and narrative information used to communicate the intended lesson content.		
7.	Number of students in the class. **(3)**		X
8.	Nature of the learning environment and expectations of the students prior to implementation of the lesson.		
9.	Philosophical orientation of the teacher regarding direct instruction. **(2)**		
10.	Nature (socio-economic status) of the majority of students.		X
11.	Educational expectations the teacher holds for the students.		
12.	Teacher effectiveness in managing a good learning environment.		
13.	Teacher motivation levels.		

	(D statements continued)	P	O
14.	Teacher's "comfort" or self-efficacy as it relates to teaching the materials.		
15.	Incentives for providing "excellent" instruction (from administration, social context, etc.).		
16.	Teacher preparation time.		
17.	Number of classes taught (other responsibilities).		X
18.	Amount of contact with class (number of class sessions per week).		X
19.	Availability of sufficient facilities and equipment.		
20.	Student characteristics (special needs students).		X
21.	Teacher experience level.		X
22.	Single-grade versus multi-grade classes.		X
23.	Support from other faculty/administrators (P.E. and non-P.E. faculty).		
E.	**Other comments related to appropriately stating the purpose and objectives for the measurement of teacher implementation fidelity**		
1.	It is important to keep in mind that the terms "implementation fidelity" and "instructional competence" are not synonymous. In fact, there are times that the "best" teachers, with the best ideas, are lowest in implementation fidelity.	X	X

Appendix D

Development of the Statement of Purpose and Objectives for the Instrument

Part A: Initial draft of the purpose and objectives for the instrument.

Draft Purpose statement

The purpose of the instrument is to assess the degree to which physical education K-2 lessons can be implemented as intended by the program developers so that the efficiency of the lesson and student achievement can be improved.

Draft Objectives

1. Assessing the implementation fidelity of nine lesson components as they are written.
2. Assessing the implementation fidelity of the time as it is allocated in the lesson components.
3. Assessing the implementation fidelity of different kinds of lesson content (fitness, activity-related knowledge, motor skills, and personal/social skills).
4. Assessing implementation fidelity by varying teacher characteristics (race, sex, experience, educational level, etc.).
5. Assessing the effectiveness of the lesson in terms of student achievement.
6. Assessing the effectiveness of the inservice training.
7. Refining the inservice training and/or instructional treatments related to student achievement.
8. Assessing the implementation fidelity in varying contexts. The lesson objectives might be implemented in different size of schools, different size classes, etc.

Part B: Revised draft of the purpose and objectives for the instrument.

Statements of purpose:

MIT member # 1

1. The purpose of the instrument is to measure the degree to which physical education teachers implement K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project as intended by program developers.

MIT member # 2

2. The purpose of the instrument is to assess the degree to which physical education lessons for grades K through 2 are implemented as they were intended by program developers.

MIT member # 3

3. The purpose of the instrument is to measure the degree to which MI-EPEC K-2 physical education lessons are implemented in accordance with the intentions of the lesson developers, evaluate the effectiveness of the lessons in regard to student achievement and identify modifications that, if implemented, will improve student achievement.

Statements of objectives:

MIT member # 1

1. Implement all of the components prescribed in the lesson.
2. Use the recommended time allocations associated with each portion of the lesson.
3. Use the teacher behaviors described in each component of the lesson.
4. Use the equipment/facilities necessary to implement the lesson as intended.

MIT member # 2

1. To assess the degree to which teachers implement each of the nine components that appear in the lessons as they were intended by lesson developers.
2. To assess the degree to which teachers implement each lesson component in the time allocated by lesson developers.
3. To assess the degree to which teachers implement lesson content representative of different domains (fitness, activity-related knowledge, motor skill, personal/social skill) in the way in which it was intended by lesson developers.
4. To create an instrument and corresponding instructions that can be used by professional physical educators with a high degree of reliability without formal inservice.
5. To create an instrument that can be applied to commercial lessons to assess the degree to which they comply to the professional literature on effective instruction (or the nine components that appear in the instrument) and are important elements of effective instruction.

MIT member # 3

1. To measure the implementation fidelity of the nine components contained in each MI-EPEC lesson.
   - 1.1 To measure the implementation fidelity of each lesson component regarding appropriate content.
   - 1.2 To measure the implementation fidelity of each lesson component regarding appropriate time allocated.

2. To measure the implementation fidelity of four lesson content areas (i.e. fitness, activity-related knowledge, motor skills, and personal/social skills).
3. To assess implementation fidelity by varying teacher characteristics (race, sex, experience, educational level, etc.).
4. To evaluate the effectiveness of the lesson in terms of student achievement.
5. To evaluate the effectiveness of the inservice training.
6. To identify the inservice training and/or instructional treatments related to student achievement.
7. To identify contextual variables related to implementation fidelity.

Part C: Final draft of the purpose and objectives for the instrument.

Approved purpose statement:

The purpose of the instrument is to measure the degree to which physical education teachers implement K-2 lessons developed by the Michigan Exemplary Physical Education Curriculum Project as written by lesson developers.

Approved objectives statements:

1. To measure the degree to which teachers implement each of the nine components that appear in the lessons as they were written by lesson developers.
2. To document the degree to which teachers implement each lesson component in accordance with the time allocated by lesson developers.
3. To provide a valid instrument for individuals who wish to document relationships that may exist between teacher implementation behavior emanating from use of MI-EPEC K-2 lessons and, for example, inservice training devoted to use of the lessons or student achievement of lesson objectives.

Appendix E

Data Collection Forms for Step 2: Creating a Table of Test Specifications

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.

C&E: Curriculum and Evaluation Specialists.

E: Evaluation Specialists.

Date_____

The table of specifications (see the attached graphic example) provides a framework for developing prototype items for measuring teacher implementation behavior. The matrix shows the components included in the MI-EPEC K-2 lesson on one dimension and information related to teacher behavior on the other. The teacher behavior side of the matrix is designed to lead to the specification of items suitable for assessing teacher implementation of the lessons.

Use of this framework will assure that relevant components of lesson implementation behavior will be evaluated. The following form provides a place for you to respond to each of the nine lesson components. For each component please write the purpose for measuring this lesson component as you understand it and the indicators you would accept as evidence of teachers appropriately implementing this component of a lesson.

The design criteria used to construct and evaluate the lessons are provided to help you complete this important task. The information you provide will be converted into prototype items for you to evaluate at a later step in the instrument development process.

Thank you for your help with this important task.

Please return your responses to Hasan by _____.

Investigator:
Hasan Al-Tawil

Graphic example of the table of test specifications

#	Content		Teacher Behavior	
	Lesson Component	**Purpose of Component**	**Achievement of Purpose (Acceptable Behavioral Indicators)**	**Prototype Items**
1.	Equipment/ material			
2.	Gym setup			
3.	Preparation of students			
4.	Explanation/ Demonstration			
5.	Transition from explanation/demon-stration to practice			
6.	Practice			
7.	Lesson summary			
8.	Review			
9.	Homework			

Step Two:
Table of Test Specifications
(Data Collection Form)

1. Equipment and materials component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

2. Gym setup component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

3. Preparation of students component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

4. Explanation/demonstration component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

5. Transition from explanation/demonstration to practice component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

6. Practice component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

7. Lesson summary component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

8. Review component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

9. Homework component

A. Purpose of this lesson component:

B. Indicators you would accept as evidence of appropriate implementation of this component:

Appendix F

Step 6: Assessing Content Validity of the Instrument

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.

C&E: Curriculum and Evaluation Specialists.

E: Evaluation Specialists.

Date_____

This step in the instrumentation process is designed to describe the content validity of the instrument. Content validity will be described in terms of the extent to which members of the MIT believe the instrument:

1. represents all components of the MI-EPEC lessons.
2. contains items that appropriately represent important teacher implementation behaviors for each component of the instrument.

Please read the attached instrument and circle the response that represents your expert opinion. If your response is "No" please identify the changes necessary to allow you to respond "Yes" in a subsequent rating.

If you have any questions regarding this step in the instrument development process please contact the investigator of the study ASAP.

Thank you for your cooperation

Investigator:
Hasan Al-Tawil

Content Appropriateness of the Components Included in the Instrument

Name _____

Please review the attached instrument and circle the response that represents your expert opinion of the appropriateness of each *component*. If your response is "No" please identify the changes necessary to allow you to respond "Yes" in a subsequent rating.

Lesson Component	Response Option	
1. Equipment/materials	Yes	No
2. Gym setup	Yes	No
3. Preparation of students	Yes	No
4. Explanation/demonstration	Yes	No
5. Transition	Yes	No
6. Practice	Yes	No
7. Review	Yes	No
8. Summary	Yes	No
9. Homework	Yes	No

Content Appropriateness of the Items Included in the Instrument

Name _____

Please review the attached instrument and circle the response that represents your expert opinion of the appropriateness of each *item.* If your response is "No" please identify the changes necessary to allow you to respond "Yes" in a subsequent rating.

1. Equipment/materials:	*(Circle Yes or No)*	
1. The *kind* of equipment/materials specified in the lesson (size, weight, color, utility or their equivalents) were provided.....	Yes	No
2. The *amount* of equipment/materials (items per student) specified in the lesson were used.....	Yes	No
3. The equipment/materials used were of good quality (condition was fully functional for the intended use)	Yes	No
4. Adapted equipment/materials, where needed for special students, were used	Yes	No

2. Gym setup:	*(Circle Yes or No)*	
1. The space (per student) used at each station for equipment, materials and students matched what was written in the lesson.....	Yes	No
2. The *number* of teaching stations described in the lesson component were used.....	Yes	No
3. The *kind* of teaching stations specified in the lesson component (distances, target sizes, available space...) were used.....	Yes	No
4. The teachers facility was of sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson	Yes	No

3. Preparation of students: *(Circle Yes or No)*

1. The students were organized as specified for this component of the lesson (seated, standing, circle ..etc..)	Yes	No
2. The objective of the lesson was clearly communicated to the students	Yes	No
3. The teacher used the lesson suggestions to clearly describe to the students why it was important to learn the lesson objective(s)	Yes	No
4. The teacher used the lesson description to connect what students were to learn with their own prior learning.....	Yes	No
5. The teacher used the key action verbs (tell, ask, restate, etc.) that were described in the lesson	Yes	No
6. The action verbs were communicated to students in the way (order, completeness, correct emphasis . .) described in the lesson.....	Yes	No
7. The "preparation" occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	Yes	No
8. Equipment/materials (if specified in this lesson component) were used as described.....	Yes	No
9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

4. Explanation/Demonstration: *(Circle Yes or No)*

Transition to explanation/demonstration

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	Yes	No
2. The transition occurred in the order described in the lesson.	Yes	No
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	Yes	No
4. The explanation of the learning intended was communicated as written.....	Yes	No
5. The cue words (or their equivalents) specified in the lesson component were fully emphasized.....	Yes	No
6. The points of emphasis specified in the lesson component (gray shaded boxes) were fully emphasized	Yes	No
7. The demonstration(s) was implemented as specified (focused on and limited to) key points of the lesson objective)	Yes	No
8. If applicable (when called for in the lesson) students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated.....	Yes	No
9. All events of the explanation/demonstration occurred in the order described in the lesson.....	Yes	No
10. The teacher was positioned where all students could participate in the explanation/demonstration.....	Yes	No

4. Explanation/Demonstration (continued): *(Circle Yes or No)*

11. The explanation/demonstration occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	Yes	No
12. Equipment/materials (if specified in this lesson component) were used as described.....	Yes	No
13. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
14. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

5. Practice: *(Circle Yes or No)*

Transition to practice

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	Yes	No
2. The transition occurred in the order described in the lesson.	Yes	No
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	Yes	No
4. The points of emphasis and cue words written in the practice activity (s) were used as described in this component of the lesson	Yes	No
5. The practice activity (s) described in the lesson were used.....	Yes	No
6. Student and teacher positioning and movement were used as specified for this practice activity	Yes	No

5. Practice (continued):	(Circle Yes or No)	
7. The number of trials suggested in the practice activity(s) were provided as described in the lesson.....	Yes	No
8. Feedback (specific to the learning task(s) and points of emphasis and connected to cue words)was delivered to students as described in this lesson component.....	Yes	No
9. The amount of feedback provided to students was as described in the lesson.....	Yes	No
10. The distribution, positioning, use and retrieval of equipment was accomplished as described in the lesson.....	Yes	No
11. When called for by the lesson, the teacher altered the lesson objective for a student(s) they achieved the criteria stated in the practice activity.	Yes	No
12. The practice occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	Yes	No
13. Equipment/materials were used as described in the lesson	Yes	No
14. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
15. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

7. Review: *(Circle Yes or No)*

Transition to review

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	Yes	No
2. The transition occurred in the order described in the lesson.	Yes	No
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	Yes	No
4. The teacher began the review by reminding the students of the lesson objectives	Yes	No
5. The teacher reviewed this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	Yes	No
6. The teacher was positioned so all students could participate in the review.....	Yes	No
7. The review occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	Yes	No
8. Equipment/materials were used as described in the lesson	Yes	No
9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

8. Lesson summary: *(Circle Yes or No)*

Transition to lesson summary

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	Yes	No
2. The transition occurred in the order described in the lesson	Yes	No
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	Yes	No
4. The teacher began the summary by reminding students of the lesson objectives.	Yes	No
5. The teacher summarized this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	Yes	No
6. The teacher was positioned so all students could participate in the summary.....	Yes	No
7. The summary occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	Yes	No
8. Equipment/materials were used as described in the lesson	Yes	No
9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

9. Homework: *(Circle Yes or No)*

Transition to homework

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	Yes	No
2. The transition occurred in the order described in the lesson.	Yes	No
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	Yes	No
4. The homework assignment was communicated as specified in the lesson.....	Yes	No
5. All materials necessary to complete the assignment were distributed as written	Yes	No
6. The homework information occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition.....)	Yes	No
7. Equipment/materials were used as described in the lesson	Yes	No
8. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	Yes	No
9. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	Yes	No

Appendix G

Inservice Agenda
Preparation for Use of the Instrument
July 2, 1998

Time	Topic	Presenter
9:00	**I. Welcome, overview, housekeeping**	Hasan
9:10	**II. Overview of *MI-EPEC* and the need for assessment of teacher implementation behavior**	Paul
	A. *MI-EPEC*	
	B. Need for evaluating teacher implementation behavior	
	C. Need for a valid instrument	
10:10	**Break**	
10:20	**III. Overview the *MI-EPEC* K-2 lessons**	Paul
	A. Lesson structure (components, what each component is designed to do)	
	B. Lesson format	
	C. Implementation procedures	
10:50	**IV. Overview the instrument**	Paul
	A. Structure (components)	
	B. Items and (development steps)	
	C. Format (organization, response format)	
	D. Overview of use (instructions, procedures of using the instrument)	
11:10	**V. Practice - Instructions for using the instrument**	Paul
	A. Familiarize yourself with the videotape system.	
	B. Carefully read the *MI-EPEC* lesson you will assess.	
	C. Review the instrument:	
	D View the videotape of the lesson	
	E. Assess teacher implementation	
12:00	**VI. Lunch**	
12:45	**V. Practice - Continued**	Paul
2: 15	**VIII. Use the instrument to evaluate a videotaped lesson**	Hasan
3:45	**IX. Terminate the session**	

Appendix H

Purpose Statements and Indicators Obtained from the MIT for Each Lesson Component

1. Equipment/Materials Component:

Purpose statements

1. To establish the degree to which the equipment and materials specified for use in the lesson:
   a. are used (kind and amount)
   b. are in good condition (clean and well maintained)
   c. includes equipment needed to service adapted students
2. To assess the degree to which equipment and materials necessary for effective and efficient instruction on the stated lesson objective were: (a) available and (b) properly used by the teacher.
3. The reason for measuring the degree to which equipment and material used in specific lessons match what was intended by the lesson writers is to provide information that can be used to assess the effects of the equipment and material used on the quality of the learning experience and students' opportunity to achieve the intended learning objective, to assess the degree to which the lesson was written well enough to support consistent implementation by a variety of teachers, and to assess the degree to which inservice is necessary and/or effective in preparing teachers to implement the lesson accordingly.

Indicators

1. The kind of equipment/materials specified in the lesson (size, weight, color, utility or its equivalent) were used.
2. The amount of equipment/materials specified in the lesson were used.
3. The equipment/materials used were clean and well maintained.
4. Adapted equipment/materials, where needed, were used.
5. Proper use of the exact type of equipment and materials specified in the written lesson.
6. Proper use of the exact amount of equipment and materials specified in the written lesson.
7. Replacing the exact type and amount of equipment and materials that were specified in the lesson only to the extent that neither the effectiveness nor efficiency of the instructional objectives (immediate or in the future) is compromised.

8. When the number of items called for by the lesson was an exact count and ample space was available, the quantity of items used exactly matched the quantity called for in the lesson.
9. When the number of items called for was written as a ratio (# per student) and ample space was available, the quantity of items used exactly matched the quantity called for in the lesson.
10. When space in the activity area was less than that called for in the lesson (e.g., the number of stations or lines had to be reduced), the quantity of items used allowed the students to engage in the activity in the way the lesson intended, with a maximum number of meaningful and safe repetitions.
11. All equipment and material was in good repair.
12. All equipment and material appeared to exactly match the following physical qualities if stipulated in the lesson:
    a. size
    b. shape
    c. weight
    d. texture
    e. compression or rebound
13. The physical qualities of equipment and/or material used allowed the students to engage in the activity in the way it was described in the lesson.

2. Gym Setup Component

Purpose statements

1. To establish the degree to which the teaching station(s) were available and were setup in accordance with the lesson:
2. To assess the degree to which:
    a. facilities necessary for effective and efficient instruction on the stated lesson objective were available.
    b. the degree to which the teacher structured and organized the physical environment (e.g., equipment, materials, facilities etc.) in a manner that is conducive to teaching/learning the stated lesson objective.
3. The purpose of this component of the lesson is to describe how to prepare the activity area so it will:
    a. maximize meaningful repetitions.
    b. engage students in a safe practice.
    c. minimize distractions.

d. position students and teacher so as to facilitate feedback and support assessment.
e. economize activity area preparation for the current day's lessons.
f. facilitate quick transitions.
g. engage students at appropriate levels of function to achieve the lesson objective.

4. The purpose of measuring the implementation fidelity is to obtain information concerning the degree to which teachers prepared the activity area as it was intended.

Indicators

1. The number of teaching stations described in the lesson can be set up in the available facility.
2. The number of teaching stations described in the lesson were used.
3. The kind teaching stations specified (dimensions) were used.
4. The teacher and student positions and movements were used as described.
5. Gym setup that is in complete accordance with that specified in the lesson.
6. Gym setup that deviates from that specified in the lesson so long as it can not be construed in any way as altering (positively or negatively; physically or psychologically) the teaching or learning of the stated lesson objectives.
7. When teachers had a sufficiently large activity area:
   a. all distances between markers, targets, and markers and targets were at distances designated in the lesson.
   b. the number of markers and targets exactly matched those listed in the lesson
   c. where and/or when distances and quantities were not designated, the activity area was set up so as to enhance a high number of safe, developmentally appropriate and meaningful repetitions.
8. When the teacher's activity area was sufficiently large, but barriers existed that prevented the activity area to be set up exactly as it was written (e.g., glass the length of one wall), or when the teacher's activity area was too small to accommodate the gym setup as it was written:
   a. the general relationships between markers, targets, and markers and targets described in the lesson was maintained.
   b. adequate space was maintained between students to support safe practice.
   c. the number of stations, lines, etc. was maximized, relative to the space available (after the first two criteria were achieved).
   d. where and/or when distances and quantities were not designated, the activity area was set up so as to enhance a high number of safe, developmentally appropriate and meaningful repetitions.

9. In all cases:

   a. all equipment and material was laid out in the activity area as the lesson described, prior to beginning the class.

3. Preparation of Students Component:

Purpose statements

1. To establish the degree to which the anticipatory sets were used as described in the lesson materials.
2. To assess the degree to which the teacher provides the students with lesson-relevant information and instruction prior to the lesson activity that is sufficient to facilitate the teaching/learning of the stated lesson objective.
3. The purpose of the component is to prepare students to learn the lesson's objective by drawing the student's attention to the objective, communicating the objective's importance, and clearly communicating what the student will attempt to learn.
4. The purpose for evaluating this component is to measure how well the teacher(s) conducted the preparation the way it was intended.

Indicators

1. The students were organized as specified for this portion of the lesson (seated, standing, circle, etc.).
2. The teacher clearly stated the objective(s) of the lesson.
3. The teacher used the lesson suggestions to clearly describe to the students <u>why</u> it was important to learn this lesson objective.
4. The teacher used the lesson description to connect what students were to learn to their own prior learning or circumstances.
5. The teacher used the lesson description to communicate to the students what they were to accomplish during the lesson.
6. Directions describing what the students were to do were clear.
7. An example of the intended learning outcome was clearly communicated to the students (definition, example, illustrations, demonstration).
8. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
9. The teacher engage in the exact student preparation that is specified in the lesson.
10. The teacher deviate from the exact student preparation that is specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) altered the intent of the lesson's preparation component.

11. Students and instructor were positioned so as to allow all students to benefit from what was said and shown.
12. The preparation occurred in correct order relative to other components (e.g., transitions).
13. The instructor used all equipment and material that was called for in the preparation.
14. The instructor used all equipment and material in the way and in the order it was explained in the lesson.
15. All key action verbs (tell, ask, restate, etc.) appearing in the lesson were presented in the lesson in correct order.
16. The key points were presented without interjecting extraneous information.
17. The preparation culminated and terminated in a clear statement of the lesson objective.
18. When important actions or information was necessary to present by the instructor, but failed to appear in the written lesson, the action or information was presented in a way the lesson would have intended.
19. When unanticipated events occurred, they were dealt with quickly and attention was refocused on the preparation content.

4. Explanation/Demonstration Component:

Purpose statements

1. To establish the degree to which the explanations and demonstrations described in the lesson materials were used as intended.
2. To assess the degree to which the teacher effectively communicated to the students the essential elements of skill contained in the lesson.

Indicators

1. The students were organized as specified for this portion of the lesson (seated, standing, circle, etc.).
2. The explanation of intended learning was communicated as written.
3. The cue words (or their equivalents) specified in the lesson were used.
4. Points of emphasis written in the lesson were used.
5. Visual materials, if specified, were used .
6. The demonstration(s) was implemented as written (limited to key points of the lesson objective).
7. The demonstration(s) were accurate.
8. Demonstrations were presented in accordance with the views prescribed in the lesson materials.

9. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
10. When described, students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated.
11. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
12. The teacher adhere totally to the explanation/demonstration instructions as they were specified in the written lesson.
13. The teacher deviate from the exact elements of the explanation/demonstration that were specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) alters the intent of the lesson's preparation component.
14. The explanation/demonstration occurred in proper sequence, relative to other lesson components.
15. All events occurred in the order described in the lesson.
16. The explanation occurred in a position relative to the students so all students benefited equally from the presentation.
17. The teacher utilized the kind of material and equipment called for by the lesson.
18. The instructor used all equipment and material in the way and in the order it was explained in the lesson.
19. The key points were presented without interjecting extraneous information.
20. The explanation occurred separate from the demonstration.
21. When important actions or information was necessary to present by the instructor, but failed to appear in the written lesson, the action or information was presented in a way the lesson would have intended.
22. When unanticipated events occurred, they were dealt with quickly and attention was refocused on the explanation/demonstration.
23. The demonstration occurred immediately after the explanation.
24. The instructor connected cue words to the explanation and demonstration, as described in the lesson.
25. The demonstrator was proficient.
26. The demonstrations occurred where all students could benefit from it equally.
27. When appropriate, students were shown the demonstration from various angles.
28. No more or less was demonstration beyond what was described in the lesson.

5. Transition from Explanation/ Demonstration to Practice Component:

Purpose statements

1. To establish the degree to which the transitions(s) described in the lessons were used.
2. To assess the degree to which the teacher provided an efficient and logical transition from the "teacher-centered" activity of explanation and demonstration to the "student-centered" activity of practice.

Indicators

1. The transition was implemented as described (words and organization).
2. The student groups called for in the transition were implemented.
3. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
4. The teacher adhere totally to the transitions from explanation/demonstration section as they were specified in the lesson.
5. The teacher deviate from the exact elements of the transitions in the explanation/demonstration specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) altered the intent of the lesson's preparation component.
6. Transitions result in placement of equipment, material and students in the manner described in the lesson.
7. The transitions occurred quickly, with no extraneous information or discourse from the teacher that slowed the process.
8. All parts of the transition occurred in the order described in the lesson.
9. Distractions were dealt with quickly.
10. All appropriate actions taken during the transition that was omitted from the lesson occurred in a reasonably efficient manner.

6. Practice Component

Purpose statements

1. To establish the degree to which the practice activities described in the lessons were used.
2. To assess the degree to which the teacher enabled maximum skill learning and development by providing all students with a sufficient amount of practice time. In addition, to assess the degree to which the teacher was capable of effectively managing the class during this somewhat unstructured practice time. Finally, to assess the degree to which the teacher was capable of accurately observing and analyzing the students'

practice performance and providing them with appropriate (in terms of both content and psychological needs of the student) feedback.

3. For students to:
   a. engage in an ample number of repetitions focused on the learning objective.
   b. maximize the quality of feedback students receive, so as to enhance their opportunity to learn.
   c. provide the opportunity for all students to learn.
   d. provide a safe practice.

Indicators

1. The key points of the practice activity(s) described in the lesson were used as written.
2. The practice activity(s) were related to the lesson objective(s) as described in the lesson.
3. Students were organized as specified for this portion of the lesson (seated, standing, circles, squads, etc.).
4. The cue words (or equivalents) of the practice activity(s) described in the lesson were used as written.
5. The number of trials suggested in the practice activity(s) described in the lesson were provided.
6. The quality of feedback provided to students was delivered as written (specific to the learning task of emphasis, connected to cue words and key points).
7. Errors in performance were corrected.
8. The amount of feedback provided to students was delivered as written.
9. Points of emphasis highlighted in the lesson were emphasized as written.
10. The distribution, positioning, use and retrieval of equipment was accomplished as written.
11. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
12. The teacher adhere totally to the practice component as it was specified in the written lesson.
13. The teacher deviate from the exact elements of the practice components specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) altered the intent of the lesson's preparation component.
14. The teacher effectively observed student performance, assesse this performance, identified the existence of any problems in this performance and provide accurate, constructive, understandable, positive feedback to all students in the class.
15. The teacher maintained an environment that is conducive to teaching and learning throughout the practice component

16. The teacher provided adequate practice time for all students in the class.
17. The teacher provides adequate supervision of all students in the class.
18. Practice began immediately after the explanation.
19. Students are engaged in the way intended by the lesson.
20. The teacher moveed among and observed all students.
21. The teacher spoke to all students concerning their performance relative to the lesson objective.
22. The teacher dealt with disruptions quickly, refocusing the students back on the lesson objective.
23. The teacher dealt with disruptions quickly, refocusing the students back on practicing the way the lesson is intended.
24. The teacher did not distract the students from the lesson objective.
25. When called for by the lesson, the teacher altered the student's objective when, and only when, they have the criteria.

7. Review Component

Purpose statements

1. To establish the degree to which the review described in the lessons was used.
2. To assess the degree to which the teacher reminded and reinforced the primary purpose of the lesson and the extent to which this was accomplished in a manner that would reinforce the essence of the lesson so as to facilitate the students' remembering the purpose of the lesson activities.

Indicators

1. The student organization called for in the review was implemented.
2. The teacher provided closure to this portion of the lesson as written (using cue words, key points and directions for how to improve their performance).
3. The teacher redirected attention to the intended outcomes (lesson objective(s)) of this portion of the lesson.
4. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
5. The teacher adhered totally to the review component as it was specified in the written lesson.
6. The teacher deviated from the exact elements of the review component specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) altered the intent of the lesson's preparation component.
7. The students were positioned where they all could benefit equally from the review.

8. The teacher began by reminding them of the lesson objectives.
9. The teacher reviewed the key points of each lesson objective.
10. The teacher presented all information in the order it was intended.
11. The teacher did not add any extraneous information before, during or after the review.
12. The teacher used the equipment and/or material called for by the review.
13. The teacher used the equipment and/or material called for by the review appropriately.

8. Summary Component

Purpose statements

1. To establish the degree to which the summary described in the lessons was used.
2. To assess the degree to which the teacher appropriately accentuated and reiterated the essential elements contained in the lesson.
3. To remind and refocus students of the essential parts of what they practiced today.

Indicators:

1. The student organization called for in the summary was implemented.
2. The teacher provided closure to this portion of the lesson as written (using cue words, key points and directions for how to improve their performance).
3. The teacher redirected attention to the intended outcomes (lesson objective(s)) of this portion of the lesson.
4. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
5. The teacher adhered totally to the lesson summary as it was specified in the written lesson.
6. The teacher deviated from the exact elements of the lesson summary specified in the written lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) alterd the intent of the lesson's summary component.
7. The students were positioned where they all could benefit equally from the summary.
8. The teacher began by reminding them of the lesson objectives.
9. The teacher reviewed the key points of each lesson objective.
10. The teacher presented all information in the order it was intended.
11. The teacher did not add any extraneous information before, during or after the summary.
12. The teacher used the equipment and/or material called for by the lesson.
13. The teacher used the equipment and/or material called for by the lesson appropriately.

9. Homework Component

Purpose statements

1. To establish the degree to which the homework intervention described in the lessons was implemented as intended.
2. To assess the degree to which the teacher assigned the students appropriate "out-of-class" activities that would reinforce the primary purpose of the lesson.

Indicators

1. The student organization called for in the summary was implemented.
2. The nature of the homework assignment was communicated as written .
3. All materials necessary to complete the assignment were distributed as written.
4. The amount of time consumed in this component of the lesson matched the amount described in the lesson.
5. The teacher adhered totally to the homework assignment specified in the lesson.
6. The teacher deviated from the exact elements of the homework assignment specified in the lesson only to the extent that such deviation in no way (positively or negatively; physically or psychologically) altered the intent of the lesson's homework component.
7. Homework is assigned.
8. All the homework is assigned (versus assigning it for only one objective if the lesson calls for both objectives).
9. The homework is assigned at the intensity and duration called for by the lesson.
10. No distracting information is presented.

10. General

Purpose statements

1. To establish the degree to which administrative, management or external disruptions influenced implementation of the lesson.

Indicators

1. There were no significant interruptions to the class (fire drill, administrative issues).
2. The teacher was an effective disciplinarian.
3. The teacher was an effective manager (student organization, teaching and practice time).

Appendix I

Step 4: Prepare Instructions for Using the Instrument

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.

C&E: Curriculum and Evaluation Specialists.

E: Evaluation Specialists.

Date: _____

Information contained in the statement of purpose and objectives for the instrument, the table of test specifications, and the revised item pool were used to write the introduction, directions for use and to establish the instrument's format.

A review draft of the instructions for the use of the instrument is attached (along with a revised instrument) for your review. Please read and react to the appropriateness of the instructions and suggest changes you believe necessary for the instrument to achieve its intended purpose. Where changes (additions, deletions, alterations) are needed, please provide a rationale for your suggestions. Write your comments directly on the instrument. Attach additional pages as needed to clearly articulate your revisions.

Thank you for your cooperation

Investigator:
Hasan Al-Tawil

Instructions for Use
Teacher Implementation Evaluation Instrument (TIEI)

Introduction

These instructions describe the purpose of the instrument, overview its general structure, describe the response format and suggest procedures for its use. Scoring, analyzing, interpreting and reporting results of the measurement process go beyond the purpose of this stage of instrument development.

Purpose of the Instrument

This instrument is designed to measure the degree to which teachers use K-2 physical education lessons as written by the MI-EPEC lesson developers.

Overview

The MI-EPEC K-2 lessons typically focus instruction on two program objectives taught during a 30 minute instructional period. Full implementation of the lessons requires 10 areas of teacher implementation behavior. Each of these areas, referred to as lesson components, is quantified by reviewer responses to a number of assessment items. Each item is an indicator of how fully a lesson component was implemented. Accordingly, reviewer responses to the items are measured using a Likert scale.

Response Format

The response options for each item use the following format: Strongly Agree (SA), Agree, Neither Agree or Disagree, Disagree and Strongly Disagree (SD). The response options also provide reviewers with a Not Applicable (NA) alternative. This is needed for instances where an item is present but the lesson does not call for the teacher behavior represented by that item. The SA - SD response format allows observers to quantify the degree to which each item was met during implementation. The rating scale is as follows:

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Not Applicable
1	2	3	4	5	NA

1. This item was *not implemented.*
2. The item was implemented to a low degree. (There were *substantial changes* in implementation that *substantially hindered* (functionally altered instruction) from what is written in the lesson.)

3. The item was implemented to a moderate degree. (There were *changes* in implementation that hindered (functionally altered instruction) what is written in the lesson.)

4. The item was implemented to a high degree. (There were *a few changes* in implementation but they *did not hinder* (or depart functionally) what is written in the lesson.)

5. The item was *fully implemented* as written in the lesson.

NA This item was not called for in the lesson.

A space is also provided to note relevant comments that would help interpret your responses to each item. These notes may be helpful at a later time to improve the quality of the instrument.

Measurement Procedures:

Observations necessary to document teacher implementation of the lessons are intended to be obtained from a videotape. To maximize the validity of the assessment, the following procedures should be used:

1. Familiarize yourself with the video player. Know:
   a. How to turn the machine on and off.
   b. How to insert, play, stop, rewind, forward and eject the tape (directly or with remote).
   c. How to control the volume.
   d. How to use the headphones, if desired.

2. Read the MI-EPEC lesson.
   Carefully read the lesson you will assess. Careful reading should cause you feel as though you are prepared to teach the lesson. It may be helpful to make personal notes similar to what you would do if you were going to teach the lesson.

3. Review the instrument.
   a. Instructions
   b. Items and response format

4. Review the videotaped lesson.
   View the lesson (or lesson portion you are assigned to assess) before conducting your assessment. Look for teacher implementation of the lesson components; seek to identify transitions from component to component.

5. Assess implementation.

Carefully complete the following steps:

1. Complete the "Evaluator Implementation Sheet."
2. Equipment/materials
   a Review this component of the written lesson.
   b. View the videotape and complete the items of this component of the instrument.
3. Replicate Step 2 above for each of the following components:
   - gym setup
   - preparation of students
   - explanation/demonstration
   - transition
   - practice
   - review
   - summary
   - homework
   - general
4. Return the results of your observations to _____.

Evaluator Information Sheet*

Evaluator Name: .....

Date of the Assessment: .....

Videotape #: .....

Grade of Lesson Assessed: ..... *Lesson Number:* .....

Lesson Objective(s): .....

Date of Lesson Implementation: .....

Teacher Identification Number: .....

Thank you for your help with this evaluation of teacher implementation of the MI-EPEC K-2 lessons.

---

* Be sure to attach all completed assessment forms for each component of the lesson to this cover sheet.

Appendix J

Teacher Implementation Evaluation Instrument (TIEI)
Instructions For Use

Introduction:

These instructions describe the purpose of the instrument, overview its general structure, describe the response format and specify procedures for its use. Scoring, analyzing, interpreting and reporting results of the measurement process go beyond the purpose of this stage of instrument development.

Purpose of the Instrument

This instrument is designed to measure the degree to which teachers implement the K-2 physical education lessons as they were written by MI-EPEC lesson developers.

Overview

Nine* areas of teacher behavior are needed to correctly implement the K-2 lessons as they were written by MI-EPEC lesson developers. Each of these areas, referred to as lesson components (equipment/materials, gym setup, preparation of the students, explanation/demonstration, transition, practice, review, summary, homework), is quantified by reviewer responses to a number of assessment items. Each item is an indicator of how fully a lesson component was implemented. Evaluator responses to the instrument's items are recorded using the response format described below.

Response Format

The response options for each item use the following format: Strongly Agree (SA), Agree, Neither Agree or Disagree, Disagree and Strongly Disagree (SD). The SA - SD response format allows observers to quantify the degree to which each item was met during implementation. The specific meaning of each point on the rating scale is as follows:

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Not Applicable
1	2	3	4	5	NA

5. The item was *fully implemented as written* in the lesson.

4. The item was implemented to a *high* degree

   There were changes in implementation but they *did not alter* instruction functionally from what is written in the lesson (i.e., implementation *was not functionally different* from what is written).

3. The item was implemented to a *moderate* degree

   There were changes in implementation that *altered* instruction from what is written in the lesson (i.e., implementation *was functionally different* from what is written).

---

* Nine lesson components, as listed above, are included in the instrument. Transitions, however, are integrated into the rating scale for all components except equipment/materials and gym setup.

2. The item was implemented to a *low* degree

   There were changes in implementation that *substantially altered* instruction from what is written in the lesson.(implementation was *functionally very different* from what is written).

1. This item was *not implemented.*

Two other response options are also provided for users of the instrument. The "Not Applicable" (NA) response is appropriate where an item is present but the written lesson does not call for the teacher behavior represented by that item. The "Not Ratable" (NR) response is appropriate where the videotape was insufficient to judge the degree of implementation.

NA This item was not called for in the lesson.

NR The videotape record was insufficient to judge the degree of implementation.

A space is also provided to note relevant comments that would help interpret your responses to each item. These notes may be helpful to improve the quality of the instrument in subsequent editions.

Measurement Procedures:

To appropriately use the instrument it is necessary to have a videotape of teacher implementation behavior (including audible sound), a written copy of the MI-EPEC lesson and a videotape playing system. To maximize the reliability and validity of the assessment, please use the following procedures:

1. Familiarize yourself with your videotape playing system. Know:
   a. How to turn the machines on and off.
   b. How to insert, play, stop, rewind, forward and eject the tape (directly or with remote).
   c. How to control the volume.
2. Carefully read the MI-EPEC lesson you will assess. Careful reading should cause you to feel as though you are prepared to teach the lesson yourself. It may be helpful to make personal notes on the written lesson.
3. Review the instrument:
   a. Instructions
   b. Items and response format
4 View the lesson (or lesson portion you are assigned to assess) *before* conducting your assessment. Identify each lesson component and seek to identify the transitions from component to component.
5. Assess teacher implementation of the MI-EPEC lesson. Carefully complete the following steps.
   a. Complete the "Evaluator Implementation Sheet."
   b. Assess the first component of the lesson (equipment and materials) using the following procedure:
      1) Review this component of the written lesson.
      2) View the videotape and complete the items of this component of the instrument.

c. Repeat this procedure for the remaining components (gym setup, preparation of students . . .)

Note:
Please assess teacher implementation of the MI-EPEC lesson as it is written!!! Do not assess the quality of instruction delivered by the teacher.

High ratings of instruction (vs. implementation of the lesson), in combination with observed changes in student behavior on the lesson's objectives, could lead MI-EPEC developers to falsely conclude that the lesson was effective when in truth it was the alternative instruction substituted by the teacher that was responsible for the obtained effects.

6. Return the results of your observations to _____.

Evaluator Information Sheet*

Evaluator Name: .....

Date of the Assessment: .....

Grade of Lesson Assessed: ..... *Lesson Number:* .....

Thank you for your help with this assessment of teacher implementation of the MI-EPEC K-2 Lessons.

---

* Be sure to attach all completed assessment forms to this cover sheet.

Teacher Implementation Evaluation Instrument (TIEI)

1. Equipment/materials:

	Strongly Disagree				Strongly Agree	
1. The *kind* of equipment/materials specified in the lesson (size, weight, color, utility or their equivalents) were provided.....	1	2	3	4	5	NA* NR**
2. The *amount* of equipment/materials (items per student) specified in the lesson were used.....	1	2	3	4	5	NA NR
3. The equipment/materials used were of good quality (condition was fully functional for the intended use).....	1	2	3	4	5	NA NR
4. Adapted equipment/materials, where needed for special students, were used	1	2	3	4	5	NA NR

2. Gym setup:

1. The space (per student) used at each station for equipment, materials and students matched what was written in the lesson	1	2	3	4	5	NA NR
2. The number of teaching stations (areas where students participate interdependently on a learning task) described in the lesson were used	1	2	3	4	5	NA NR
3. The *kind* of teaching stations specified in the lesson component were used.....	1	2	3	4	5	NA NR
4. The facility was of sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson	1	2	3	4	5	NA NR

---

* NA = Not Applicable: An example would be the presence of an item that has no corresponding element in the written lesson.

** NR = Not Ratable: An example would be where the videotape was insufficient to judge the degree of implementation.

3. Preparation of students:

	Strongly Disagree				Strongly Agree	
1. The students were organized as specified for this component of the lesson (seated, standing, circle, etc.)	1	2	3	4	5	NA NR
2. The objective of the lesson was clearly communicated to the students	1	2	3	4	5	NA NR
3. The teacher used the lesson suggestions to clearly describe to the students why it was important to learn the lesson objective(s)	1	2	3	4	5	NA NR
4. The teacher used the lesson description to connect what students were to learn with their own prior learning.....	1	2	3	4	5	NA NR
5. The teacher used the key action verbs (tell, ask, restate, etc.) that were written in the lesson	1	2	3	4	5	NA NR
6. The action verbs were communicated to students in the manner (order, completeness, correct emphasis...) described in the lesson.....	1	2	3	4	5	NA NR
7. The "preparation" occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
8. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

4. Explanation/Demonstration:

	Strongly Disagree				Strongly Agree	
1. The explanation of the learning intended was communicated as written.....	1	2	3	4	5	NA NR
2. The cue words (or their equivalents) specified in the lesson component were fully emphasized....	1	2	3	4	5	NA NR
3. The points of emphasis specified in the lesson component (gray shaded boxes) were fully emphasized.....	1	2	3	4	5	NA NR
4. The demonstration(s) was implemented as specified (focused on and limited to key points of the lesson objective)	1	2	3	4	5	NA NR
5. When written in the lesson, students had the opportunity to demonstrate their understanding of each cue word used	1	2	3	4	5	NA NR
6. All events of the explanation/demonstration occurred in the order described in the lesson.....	1	2	3	4	5	NA NR
7. The teacher was positioned where all students could participate in the explanation/demonstration.....	1	2	3	4	5	NA NR
8. The explanation/demonstration occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
9. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA NR

4. Explanation/Demonstration (continued):

	Strongly Disagree				Strongly Agree	
Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.						
10. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
11. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR
Note: Use the following three items if there is a transition included in this component of the lesson - otherwise ignore these three items.						
12. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
13. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
14. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR

5. Practice:

	Strongly Disagree				Strongly Agree	
1. The points of emphasis and cue words written in the practice activity(s) were used as described in this component of the lesson.....	1	2	3	4	5	NA NR
2. The practice activity(s) described in the lesson were used.....	1	2	3	4	5	NA NR
3. Student and teacher positioning and movement were used as specified for this practice activity	1	2	3	4	5	NA NR
4. The number of trials suggested in the practice activity(s) were provided as described in the lesson.....	1	2	3	4	5	NA NR
5. Feedback (specific to the learning task(s) and points of emphasis and connected to cue words) was delivered to students as described in this lesson component..	1	2	3	4	5	NA NR
6. The amount of feedback provided to students was as described in the lesson.....	1	2	3	4	5	NA NR
7. The distribution, positioning, use and retrieval of equipment/materials was accomplished as described in the lesson.....	1	2	3	4	5	NA NR
8. When called for by the lesson, the teacher altered the lesson objective for a student(s) who achieved the criteria stated in the practice activity.....	1	2	3	4	5	NA NR
9. The practice occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
10. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

5. Practice (continued):

	Strongly Disagree				Strongly Agree	
Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.						
11. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
12. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR
Note: Use the following three items if there is a transition included in this component of the lesson - otherwise ignore these three items.						
13. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
14. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
15. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR

6. Review:

	Strongly Disagree				Strongly Agree	
1. The teacher began the review by reminding the students of the lesson objectives	1	2	3	4	5	NA NR
2. The teacher reviewed this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA NR
3. The teacher was positioned so all students could participate in the review.....	1	2	3	4	5	NA NR
4. The review occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
5. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

6. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
7. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

Note: Use the following three items if there is a transition included in this component of the lesson - otherwise ignore these three items.

8. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
9. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR

6. Review (continued):

	Strongly Disagree				Strongly Agree	
10. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR

7. Lesson summary:

1. The teacher began the summary by reminding students of the lesson objectives.	1	2	3	4	5	NA NR
2. The teacher summarized this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA NR
3. The teacher was positioned so all students could participate in the summary.....	1	2	3	4	5	NA NR
4. The summary occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
5. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

6. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
7. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

Note: Use the following three items if there is a transition included in this component of the lesson - otherwise ignore these three items.

7. Lesson summary (continued):

	Strongly Disagree				Strongly Agree	
8. The transition process (words, strategies, actions) was implemented as described in this component of the lesson.....	1	2	3	4	5	NA NR
9. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
10. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR

8. Homework:

1. The homework assignment was communicated as specified in the lesson.....	1	2	3	4	5	NA NR
2. All materials necessary to complete the assignment were distributed as written	1	2	3	4	5	NA NR
3. The homework information occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition...)	1	2	3	4	5	NA NR
4. Equipment/materials were used as described in the lesson.....	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

5. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
6. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

Note: Use the following three items if there is a transition included in this component of the lesson - otherwise ignore these three items.

8. Homework: (continued):

	Strongly Disagree				Strongly Agree	
7. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
8. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
9. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR

Appendix K

Second Edition of Instrument
Obtained as a Result of Completion of Steps 1-4

Instructions for Using the Instrument

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.

C&E: Curriculum and Evaluation Specialists.

E: Evaluation Specialists.

Date: _____

Information contained in the statement of purpose and objectives for the instrument, the table of test specifications, and the revised item pool were used to write the introduction, directions for use and to establish the instrument's format.

A review draft of the instructions for the use of the instrument is attached (along with a revised instrument) for your review. Please read and react to the appropriateness of the instructions and suggest changes you believe necessary for the instrument to achieve its intended purpose. Where changes (additions, deletions, alterations) are needed, please provide a rationale for your suggestions. Write your comments directly on the instrument. Attach additional pages as needed to clearly articulate your revisions.

Thank you for your cooperation

Investigator:
Hasan Al-Tawil

Instructions for Use

Teacher Implementation Evaluation Instrument (TIEI)

Introduction:

These instructions describe the purpose of the instrument, overview its general structure, describe the response format and specifies procedures for its use. Scoring, analyzing, interpreting and reporting results of the measurement process go beyond the purpose of this stage of instrument development.

Purpose of the Instrument

This instrument is designed to measure the degree to which teachers use K-2 physical education lessons as written by the MI-EPEC lesson developers.

Overview

Implementation of the MI-EPEC lessons as they are written requires nine areas of teacher behavior. Each of these areas, referred to as lesson components (equipment/materials, gym setup, preparation of the students, explanation/ demonstration, transitions, practice, review, summary, homework), is quantified by reviewer responses to a number of assessment items. Each item is an indicator of how fully a lesson component was implemented. Reviewer responses to the items are measured using a modified Likert scale.

Response Format

The response options for each item use the following format: Strongly Agree (SA), Agree, Neither Agree or Disagree, Disagree and Strongly Disagree (SD). The response options also provide reviewers with a Not Applicable (NA) alternative. The NA response is needed in instances where an item is present but the written lesson does not call for the teacher behavior represented by that item. The SA - SD response format allows observers to quantify the degree to which each item was met during implementation. The rating scale is as follows:

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Not Applicable
1	2	3	4	5	NA

1. This item was *not implemented.*
2. The item was implemented to a *low* degree.

   There were changes in implementation that *substantially altered* instruction from what is written in the lesson. (implementation was functionally very different from what is written).
3. The item was implemented to a *moderate* degree.

   There were changes in implementation that *altered* instruction from what is written in the lesson (implementation was different from what is written).
4. The item was implemented to a *high* degree.

   There were changes in implementation but they *did not alter* instruction functionally from what is written in the lesson (implementation was not functionally different from what is written).
5. The item was *fully implemented* as written in the lesson.

NA This item was not called for in the lesson.

A space is also provided to note relevant comments that would help interpret your responses to each item. These notes may be helpful at a later time to improve the quality of the instrument.

Measurement Procedures:

Observations necessary to document teacher implementation of the lessons are intended to be obtained from a videotape. To maximize the validity of the assessment, the following procedures should be used:

1. Familiarize yourself with the video player. Know:
   a. How to turn the machine on and off.
   b. How to insert, play, stop, rewind, forward and eject the tape (directly or with remote).
   c. How to control the volume.
   d. How to use the headphones, if desired.
2. Read the MI-EPEC lesson.

   Carefully read the lesson you will assess. Careful reading should cause you to feel as though you are prepared to teach the lesson. It may be helpful to make personal notes similar to what you would do if you were going to teach the lesson.
3. Review the instrument.
   a. Instructions
   b. Items and response format
4. Review the videotaped lesson.

   View the lesson (or lesson portion you are assigned to assess) before conducting your assessment. Look for teacher implementation of the lesson components; seek to identify transitions from component to component.
5. Assess implementation.

   Carefully complete the following steps:
   1. Complete the "Evaluator Implementation Sheet."
   2. Equipment/materials
      a Review this component of the written lesson.
      b. View the videotape and complete the items of this component of the instrument.
   3. Replicate Step 2 above for each of the following components
      - gym setup
      - preparation of students
      - explanation/demonstration
      - transition
      - practice
      - review
      - summary
      - homework
      - general
   4. Return the results of your observations to _____.

Evaluator Information Sheet*

Evaluator Name: .....

Date of Assessment: .....

Videotape #: .....

Grade of Lesson Assessed: ..... *Lesson Number:* .....

Lesson Objective(s): .....

Date of Lesson Implementation: .....

Teacher Identification Number: .....

Thank you for your help with this evaluation of teacher implementation of the MI-EPEC K-2 lessons.

---

* Be sure to attach all completed assessment forms for each component of the lesson to this cover sheet.

Revised Prototype Items for the Teacher Implementation Evaluation Instrument (TIEI)

1. Equipment/Materials:

	Strongly Disagree				Strongly Agree	
1. The *kind* of equipment/materials specified in the lesson (size, weight, color, utility or their equivalents) were used.....	1	2	3	4	5	NA*
2. The *amount* of equipment/materials (items per student) specified in the lesson were used.....	1	2	3	4	5	NA
3. The equipment/materials used were of good quality (condition was fully functional for the intended use)	1	2	3	4	5	NA
4. Adapted equipment/materials, where needed for special students, were used	1	2	3	4	5	NA

2. Gym Setup:

1. The space used at each station for equipment, materials and students matched what was written in the lesson	1	2	3	4	5	NA
2. The *number* of teaching stations described in the lesson component were used.....	1	2	3	4	5	NA
3. The *kind* of teaching stations specified in the lesson component (distances, target sizes, available space. . .) were used.....	1	2	3	4	5	NA
4. The teacher's facility was of sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson	1	2	3	4	5	NA

* NA = Not Applicable: An example would be the presence of an item that has no corresponding element in the written lesson.

3. Preparation of Students:

	Strongly Disagree				Strongly Agree	
1. The students were organized as specified for this component of the lesson (seated, standing, circle, etc.)	1	2	3	4	5	NA
2. The teacher clearly communicated (statement, example, definition, illustration, demonstration) the objective(s) of the lesson (intended learning outcomes) to the students.....	1	2	3	4	5	NA
3. The teacher used the lesson suggestions to clearly describe to the students why it was important to learn the lesson objective(s)	1	2	3	4	5	NA
4. The teacher used the lesson description to connect what students were to learn with their own prior learning.....	1	2	3	4	5	NA
5. The teacher used the key action verbs (tell, ask, restate, etc.) to communicate to the students what they were expected to accomplish during the lesson.....	1	2	3	4	5	NA
6. The preparation occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA
7. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA
8. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	1	2	3	4	5	NA
9. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

4. Explanation/Demonstration:

	Strongly Disagree				Strongly Agree	
1. The students were organized as specified for this portion of the lesson component (seated, standing, circle, etc.)	1	2	3	4	5	NA
2. The explanation of the learning intended was communicated as written	1	2	3	4	5	NA
3. The cue words (or their equivalents) specified in the lesson component were fully emphasized	1	2	3	4	5	NA
4. The points of emphasis specified in the lesson component (gray shaded boxes) were fully emphasized	1	2	3	4	5	NA
5. The demonstration(s) was implemented as specified (focused on and limited to) key points of the lesson objective)	1	2	3	4	5	NA
6. If applicable (when called for in the lesson) students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated	1	2	3	4	5	NA
7. All events of the explanation/demonstration occurred in the order described in the lesson	1	2	3	4	5	NA
8. The teacher was positioned where all students could participate in the explanation/demonstration	1	2	3	4	5	NA
9. The explanation/demonstration occurred in correct order relative to other lesson components (e.g., preparation, explanation/ demonstration, transition. . .)	1	2	3	4	5	NA
10. Equipment/materials (if specified in this lesson component) were used as described	1	2	3	4	5	NA

4. Explanation/Demonstration (Continued):	Strongly Disagree				Strongly Agree	
11. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	1	2	3	4	5	NA
12. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

5. Transition:						
1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA
2. All parts of the transition occurred in the order described in the lesson.....	1	2	3	4	5	NA
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA

6. Practice:						
1. The points of emphasis and cue words written in the practice activity(s) were used as described in this component of the lesson.....	1	2	3	4	5	NA
2. The practice activity(s) described in the lesson were used.....	1	2	3	4	5	NA
3. Student and teacher positioning and movement were used as specified for this practice activity (seated, standing, circles, squads, etc.)	1	2	3	4	5	NA
4. The number of trials suggested in the practice activity(s) were provided as described in the lesson.....	1	2	3	4	5	NA

6. Practice (Continued):

	Strongly Disagree				Strongly Agree	
5. Feedback (specific to the learning task(s) and points of emphasis and connected to cue words) was delivered to students as described in this lesson component.....	1	2	3	4	5	NA
6. The amount of feedback provided to students was as described in the lesson	1	2	3	4	5	NA
7. The distribution, positioning, use and retrieval of equipment was accomplished as described in the lesson.....	1	2	3	4	5	NA
8. When called for by the lesson, the teacher altered the lesson objective for a student(s) when they achieved the criteria stated in the practice activity. .	1	2	3	4	5	NA
9. The practice occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA
10. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA
11. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	1	2	3	4	5	NA
12. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

7. Review:

1. The teacher began the review by reminding the students of the lesson objectives	1	2	3	4	5	NA

7. Review (Continued):

	Strongly Disagree				Strongly Agree	
2. The teacher reviewed this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA
3. The teacher was positioned so all students could participate in the review.....	1	2	3	4	5	NA
4. The review occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA
5. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA
6. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	1	2	3	4	5	NA
7. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

8. Lesson Summary:

1. The teacher began the summary by reminding students of the lesson objectives.....	1	2	3	4	5	NA
2. The teacher summarized this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA
3. The teacher was positioned so all students could participate in the summary.....	1	2	3	4	5	NA

8. Lesson Summary (Continued):

	Strongly Disagree				Strongly Agree	
4. The summary occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA
5. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA
6. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component.....	1	2	3	4	5	NA
7. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

9. Homework:

1. The specified homework assignment was communicated as specified in this lesson component	1	2	3	4	5	NA
2. All materials necessary to complete the assignment were distributed as written	1	2	3	4	5	NA
3. The homework information occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA
4. Equipment/materials (if specified in the lesson component) were used as described.....	1	2	3	4	5	NA
5. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is emphasized in this lesson component.....	1	2	3	4	5	NA
6. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA

Appendix L

Third Edition of Instrument
Obtained as a Result of Completion of Step 5.1

Project staff member, please circle one: C&I C&E E

C&I: Curriculum and Instruction Specialists.

C&E: Curriculum and Evaluation Specialists.

E: Evaluation Specialists.

Date_____

Completion of step 5 involves pilot testing the instrument by rating a teacher teaching a 15 minute MI-EPEC lesson. After you rate the videotape, a meeting will be held to identify the difficulties you encountered while using the instrument and to refine the instrument so that these difficulties are resolved.

Please keep the following criteria in mind as you assess teacher implementation as recorded on the attached videotape.

1. Item consistency with purpose of each lesson component.
2. Discreteness of the behavioral description.
3. Measurability.
4. Clarity.
5. Lack of bias.
6. Other issues that may affect the validity of the instrument.

Alterations in the instructions or the instrument's format which will facilitate the appropriate use of the instrument are also requested. Again, include the rationale for the changes you suggest.

Thank you for your cooperation

Investigator:
Hasan Al-Tawil

Instructions for Use

Teacher Implementation Evaluation Instrument (TIEI)

Introduction:

These instructions describe the purpose of the instrument, overview its general structure, describe the response format and specify procedures for its use. Scoring, analyzing, interpreting and reporting results of the measurement process go beyond the purpose of this stage of instrument development.

Purpose of the Instrument

This instrument is designed to measure the degree to which teachers implement the K-2 physical education lessons as they were written by MI-EPEC lesson developers.

Overview

Nine* areas of teacher behavior are required to correctly implement the K-2 lessons as they were written by MI-EPEC lesson developers. Each of these areas, referred to as lesson components (equipment/materials, gym setup, preparation of the students, explanation/demonstration, transition, practice, review, summary, homework), is quantified by reviewer responses to a number of assessment items. Each item is an indicator of how fully a lesson component was implemented. Evaluator responses to the instrument's items are recorded using the response format described below.

Response Format

The response options for each item use the following format: Strongly Agree (SA), Agree, Neither agree or Disagree, Disagree and Strongly Disagree (SD). The SA - SD response format allows observers to quantify the degree to which each item was met during implementation. The specific meaning of each point on the rating scale is as follows:

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Not Applicable
1	2	3	4	5	NA

5. The item was *fully implemented as written* in the lesson.

4. The item was implemented to a *high* degree.

   There were changes in implementation but they *did not alter* instruction functionally from what is written in the lesson (i.e., implementation *was not functionally different* from what is written).

3. The item was implemented to a *moderate* degree.

   There were changes in implementation that *altered* instruction from what is written in the lesson (i.e., implementation *was functionally different* from what is written).

2. The item was implemented to a *low* degree.

   There were changes in implementation that *substantially altered* instruction from what is written in the lesson.(implementation was *functionally very different* from what is written).

1. This item was *not implemented.*

* Nine lesson components as listed above are included in the instrument. Transitions, however, are integrated into the rating scale as the first three items for the components other than equipment/materials and gym setup.

Two other response options are also provided for users of the instrument. A "Not Applicable" (NA) response is needed in instances where an item is present but the written lesson does not call for the teacher behavior represented by that item. A "Not Ratable" (NR) response is also needed. It is appropriate to use the NR option in instances where the videotape record was insufficient to judge the degree of implementation.

NA This item was not called for in the lesson.

NR The videotape record was insufficient to judge the degree of implementation.

A space is also provided to note relevant comments that would help interpret your responses to each item. These notes may be helpful to improve the quality of the instrument in subsequent editions.

Measurement Procedures:

To appropriately use the instrument it is necessary to have a videotape of teacher implementation behavior (including audible sound), a written copy of the MI-EPEC lesson and a videotape playing system. To maximize the reliability and validity of the assessment, please use the following procedures:

1. Familiarize yourself with your videotape playing system. Know:
   a. How to turn the machines on and off.
   b. How to insert, play, stop, rewind, forward and eject the tape (directly or with remote).
   c. How to control the volume.
2. Carefully read the MI-EPEC lesson you will assess. Careful reading should cause you to feel as though you are prepared to teach the lesson. It may be helpful to make personal notes on the written lesson.
3. Review the instrument:
   a. Instructions
   b. Items and response format
4. View the lesson (or lesson portion you are assigned to assess) *before* conducting your assessment. Identify each lesson component and seek to identify the transitions from component to component.
5. Assess teacher implementation of the MI-EPEC lesson. Carefully complete the following steps:
   a. Complete the "Evaluator Implementation Sheet."
   b. Assess the first component of the lesson (equipment and materials) using the following procedure:
      1) Review this component of the written lesson.
      2) View the videotape and complete the items of this component of the instrument.

      Note: *Assess teacher implementation of the written MI-EPEC lesson!!! Don't* assess the quality of instruction delivered by the teacher. High ratings of instruction (versus implementation of the lesson), in combination with observed changes in student behavior on the lesson's objectives, could lead MI-EPEC developers to falsely conclude that the lesson was effective when in truth it was the alternative instruction substituted by the teacher that was responsible for the obtained effects.
   c. Repeat this procedure for the remaining components (gym setup, preparation of students . . .).
6. Return the results of your observations to _____.

Evaluator Information Sheet*

Evaluator Name: .....

Date of the Assessment: .....

Grade of Lesson Assessed: ..... *Lesson Number:* .....

Lesson Objective(s): .....

.....

Thank you for your help with this assessment of teacher implementation of the MI-EPEC K-2 lessons.

---

* Be sure to attach all completed assessment forms to this cover sheet.

Revised Prototype Items for the Teacher Implementation Evaluation Instrument (TIEI)

1. Equipment/Materials:

	Strongly Disagree				Strongly Agree	
1. The *kind* of equipment/materials specified in the lesson (size, weight, color, utility or their equivalents) were provided.....	1	2	3	4	5	NA* NR**
2. The *amount* of equipment/materials (items per student) specified in the lesson were used.....	1	2	3	4	5	NA NR
3. The equipment/materials used were of good quality (condition was fully functional for the intended use)	1	2	3	4	5	NA NR
4. Adapted equipment/materials, where needed for special students, were used	1	2	3	4	5	NA NR

2. Gym Setup:

1. The space (per student) used at each station for equipment, materials and students matched what was written in the lesson	1	2	3	4	5	NA NR
2. The *number* of teaching stations described in the lesson component were used.....	1	2	3	4	5	NA NR
3. The *kind* of teaching stations specified in the lesson component (distances, target sizes, available space. . .) were used.....	1	2	3	4	5	NA NR
4. The teachers facility was of sufficient quality (size, walls, ceiling surface) to accommodate full implementation of the lesson	1	2	3	4	5	NA NR

* NA = Not Applicable: An example would be the presence of an item that has no corresponding element in the written lesson.

** NR = Not Ratable: An example would be where the videotape was insufficient to judge the degree of implementation.

3. Preparation of Students:

	Strongly Disagree				Strongly Agree	
1. The students were organized as specified for this component of the lesson (seated, standing, circle, etc.)	1	2	3	4	5	NA NR
2. The objective of the lesson was clearly communicated to the students	1	2	3	4	5	NA NR
3. The teacher used the lesson suggestions to clearly describe to the students why it was important to learn the lesson objective(s)	1	2	3	4	5	NA NR
4. The teacher used the lesson description to connect what students were to learn with their own prior learning.....	1	2	3	4	5	NA NR
5. The teacher used the key action verbs (tell, ask, restate, etc.) that were described in the lesson	1	2	3	4	5	NA NR
6. The action verbs were communicated to students in the way (order, completeness, correct emphasis . . .) described in the lesson.....	1	2	3	4	5	NA NR
7. The "preparation" occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR
8. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

4. Explanation/Demonstration:	Strongly Disagree				Strongly Agree	
Transition to explanation/demonstration						
1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
2. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR
4. The explanation of the learning intended was communicated as written.....	1	2	3	4	5	NA NR
5. The cue words (or their equivalents) specified in the lesson component were fully emphasized....	1	2	3	4	5	NA NR
6. The points of emphasis specified in the lesson component (gray shaded boxes) were fully emphasized.....	1	2	3	4	5	NA NR
7. The demonstration(s) was implemented as specified (focused on and limited to) key points of the lesson objective.....	1	2	3	4	5	NA NR
8. If applicable (when called for in the lesson) students had the opportunity to demonstrate their understanding of the intended learning for each cue communicated.....	1	2	3	4	5	NA NR
9. All events of the explanation/demonstration occurred in the order described in the lesson.....	1	2	3	4	5	NA NR
10. The teacher was positioned where all students could participate in the explanation/ demonstration.....	1	2	3	4	5	NA NR

4. Explanation/Demonstration (continued):

	Strongly Disagree				Strongly Agree	
11. The explanation/demonstration occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR
12. Equipment/materials (if specified in this lesson component) were used as described.....	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

13. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
14. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

5. Practice:

Transition to practice

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
2. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR
4. The points of emphasis and cue words written in the practice activity(s) were used as described in this component of the lesson.....	1	2	3	4	5	NA NR

5. Practice (continued):

	Strongly Disagree				Strongly Agree	
5. The practice activity(s) described in the lesson were used.....	1	2	3	4	5	NA NR
6. Student and teacher positioning and movement were used as specified for this practice activity.....	1	2	3	4	5	NA NR
7. The number of trials suggested in the practice activity(s) were provided as described in the lesson	1	2	3	4	5	NA NR
8. Feedback (specific to the learning task(s) and points of emphasis and connected to cue words) was delivered to students as described in this lesson component..	1	2	3	4	5	NA NR
9. The amount of feedback provided to students was as described in the lesson.....	1	2	3	4	5	NA NR
10. The distribution, positioning, use and retrieval of equipment was accomplished as described in the lesson.....	1	2	3	4	5	NA NR
11. When called for by the lesson, the teacher altered the lesson objective for a student(s) when they achieved the criteria stated in the practice activity...	1	2	3	4	5	NA NR
12. The practice occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR
13. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

14. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR

5. Practice (continued):

	Strongly Disagree				Strongly Agree	
15. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

7. Review:

Transition to review

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
2. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR
4. The teacher began the review by reminding the students of the lesson objectives	1	2	3	4	5	NA NR
5. The teacher reviewed this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA NR
6. The teacher was positioned so all students could participate in the review.....	1	2	3	4	5	NA NR
7. The review occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR
8. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

7. Review (Continued):

	Strongly Disagree				Strongly Agree	
9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

8. Lesson Summary:

Transition to lesson summary

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
2. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR
4. The teacher began the summary by reminding students of the lesson objectives.	1	2	3	4	5	NA NR
5. The teacher summarized this portion of the lesson as written (used key points, cue words and directions for how to improve their performance)	1	2	3	4	5	NA NR
6. The teacher was positioned so all students could participate in the summary.....	1	2	3	4	5	NA NR
7. The summary occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR

8. Lesson Summary (Continued):

	Strongly Disagree				Strongly Agree	
8. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

9. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
10. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

9. Homework:

Transition to homework

1. The transition process (words, strategies, actions) was implemented as described in this component of the lesson	1	2	3	4	5	NA NR
2. The transition occurred in the order described in the lesson.	1	2	3	4	5	NA NR
3. The transition process resulted in placement of the students, equipment and materials as described in this component of the lesson.....	1	2	3	4	5	NA NR
4. The homework assignment was communicated as specified in the lesson.....	1	2	3	4	5	NA NR
5. All materials necessary to complete the assignment were distributed as written	1	2	3	4	5	NA NR
6. The homework information occurred in correct order relative to other lesson components (e.g., preparation, explanation/demonstration, transition. . .)	1	2	3	4	5	NA NR

9. Homework (continued):

	Strongly Disagree				Strongly Agree	
7. Equipment/materials were used as described in the lesson	1	2	3	4	5	NA NR

Note: For the following two items use only the SA - SD portion of the scale. Ignore the part of the scale that refers to degree of implementation of the lesson as written.

8. The key points were presented without unnecessary comments (extraneous information) that extended beyond what is written in this lesson component...	1	2	3	4	5	NA NR
9. The teacher managed unanticipated events (distractions) by quickly refocusing attention to the intent of this component of the lesson	1	2	3	4	5	NA NR

Appendix M

Degree of Agreement Among Members of the MIT

Item	Response (%) by category							Criterion score	Agreement among numerical responses	Agreement among all responses
	1	2	3	4	5	NA	NR			
1.	0	0	0	0	100	0	0	5	1	
2.	0	33	33	0	33	0	0	2	4	
3.	0	0	0	0	100	0	0	5	1	
4.	0	0	0	0	0	100	0	NA		1
5.	0	0	0	33	67	0	0	5	2	
6.	0	100	0	0	0	0	0	2	1	
7.	67	33	0	0	0	0	0	1	2	
8.	0	0	0	33	67	0	0	5	2	
9.	33	33	33	0	0	0	0	1	3	
10.	100	0	0	0	0	0	0	1	1	
11.	33	33	33	0	0	0	0	2	3	
12.	33	33	33	0	0	0	0	2	3	
13.	0	33	67	0	0	0	0	3	2	
14.	0	67	33	0	0	0	0	2	2	
15.	0	0	0	0	100	0	0	5	1	
16.	0	0	0	0	0	100	0	NA		1
17.	0	0	67	33	0	0	0	3	2	
18.	0	0	0	0	67	33	0	5		2
19.	0	100	0	0	0	0	0	2	1	
20.	0	0	0	0	0	100	0	NA		1
21.	0	0	0	0	0	67	33	NA		1
22.	0	67	0	0	0	0	33	2		2
23.	0	0	0	0	0	0	100	NA		1
24.	0	67	33	0	0	0	0	2	2	
25.	0	67	33	0		0	0	2	2	
26.	0	0	0	0	100	0	0	5	1	
27.	0	100	0	0	0	0	0	2	1	
28.	0	33	67	0	0	0	0	2	2	
29.	0	0	0	0	0	100	0	NA		1
30.	100	0	0	0	0	0	0	1	1	
31.	100	0	0	0	0	0	0	1	1	
32.	0	33	67	0	0	0	0	3	2	
33.	0	100	0	0	0	0	0	2	1	
34.	0	100	0	0	0	0	0	2	1	
35.	67	33	0	0	0	0	0	1	2	

Item	1	2	3	4	5	NA	NR	Criterion score	Agreement among numerical responses	Agreement among all responses
36.	0	0	0	0	0	0	100	NR		1
37.	0	67	33	0	0	0	0	2	2	
38.	0	100	0	0	0	0	0	2	1	
39.	0	67	0	0	33	0	0	2	4	
40.	0	0	0	0	0	100	0	NA		1
41.	0	0	0	0	100	0	0	5	1	
42.	33	33	33	0	0	0	0	2	3	
43.	0	100	0	0	0	0	0	2	1	
44.	0	0	0	0	0	100	0	NA		1
45.	0	0	0	0	0	100	0	NA		1
46.	0	0	0	0	0	100	0	NA		1
47.	0	0	0	0	0	100	0	NA		1
48.	0	0	0	0	0	100	0	NA		1
49.	0	67	33	0	0	0	0	2	2	
50.	0	33	33	0	33	0	0	3	4	
51.	0	0	0	0	100	0	0	5	1	
52.	0	0	0	0	0	100	0	NA		1
53.	0	0	0	75	25	0	0	4	2	
54.	0	0	0	0	0	100	0	NA		1
55.	0	0	0	67	33	0	0	4	2	
56.	0	0	0	0	100	0	0	5	1	
57.	0	33	33	0	33	0	0	4	4	
58.	33	0	0	0	0	67	0	NA		2
59.	0	0	100	0	0	0	0	3	1	
60.	0	33	33	0	33	0	0	3	4	
61.	0	0	0	0	100	0	0	5	1	
62.	0	0	0	0	0	100	0	NA		1
63.	33	33	33	0	0	0	0	2	3	
64.	0	33	0	33	33	0	0	3	4	
65.	0	0	0	0	0	100	0	NA		1
66.	0	0	0	0	0	100	0	NA		1
67.	0	0	0	0	0	100	0	NA		1
68.	0	67	33	0	0	0	0	2	2	
69.	0	0	0	0	0	100	0	NA		1
70.	0	0	0	0	100	0	0	5	1	
71.	0	0	0	0	0	100	0	NA		1
72.	0	33	33	33	0	0	0	3	3	
73.	0	0	0	33	0	67	0	NA		2
74.	0	0	0	0	0	100	0	NA		1
75.	0	0	0	0	0	100	0	NA		1
76.	0	0	0	0	0	100	0	NA		1

The degree of agreement represented by the valuaes devoted to the last two columns in the table above "agreement among numerical responses" and "agreement among all reponses" as follows:

1 = Perfect agreement by all raters

2 = All ratings contained in 2 adjacent categories

3 = All ratings contained in 3 adjacent categories

4 = All ratings contained in 4 adjacent categories

5 = All ratings contained in 5 adjacent categories

Note: Because the two categorical reponses "NA" and "NR" could be scored differently in the agreement column if their order was reveresed in the table they are considered one category for the purpose of entering an agreement score in the last column.

Appendix N

Degree of Agreement Among Members of the SRT

Response (%) by category										
Item	1	2	3	4	5	NA	NR	Criterion score	Agreement among numerical responses	Agreement among all responses
1.	0	0	0	0	100	0	0	5	1	
2.	0	25	25	25	0	0	25	2		4
3.	0	0	0	0	100	0	0	5	1	
4.	0	0	0	0	0	100	0	NA		1
5.	0	50	0	0	50	0	0	5	4	
6.	0	25	25	0	50	0	0	2	4	
7.	0	0	25	0	75	0	0	1	3	
8.	0	0	50	0	50	0	0	5	3	
9.	25	25	50	0	0	0	0	1	3	
10.	25	25	25	25	0	0	0	1	4	
11.	0	0	50	50	0	0	0	2	2	
12.	0	0	25	25	50	0	0	2	3	
13.	0	0	25	50	25	0	0	3	3	
14.	0	0	25	50	25	0	0	2	3	
15.	0	0	0	0	100	0	0	5	1	
16.	0	0	0	0	0	100	0	NA		1
17.	0	25	25	50	0	0	0	3	3	
18.	0	0	0	25	75	0	0	5	2	
19.	0	50	25	25	0	0	0	2	3	
20.	25	0	25	25	25	0	0	NA	5	
21.	0	0	0	0	0	100	0	NA		1
22.	0	0	0	50	25	0	25	2		3
23.	25	0	0	0	0	0	75	NA		2
24.	0	50	25	25	0	0	0	2	3	
25.	0	25	25	0	50	0	0	2	4	
26.	0	0	25	0	75	0	0	5	3	
27.	0	25	25	25	25	0	0	2	4	
28.	0	25	50	25	0	0	0	2	3	
29.	0	0	0	0	50	50	0	NA		2
30.	25	50	25	0	0	0	0	1	3	
31.	25	25	50	0	0	0	0	1	3	
32.	25	50	0	25	0	0	0	3	4	
33.	25	0	25	25	25	0	0	2	5	
34.	0	50	25	0	25	0	0	2	4	
35.	0	100	0	0	0	0	0	1	1	

Item	1	2	3	4	5	NA	NR	Criterion score	Agreement among numerical responses	Agreement among all responses
36.	0	0	25	25	25	25	0	NR		4
37.	0	25	0	25	25	25	0	2		5
38.	0	0	25	25	0	50	0	2		3
39.	0	75	25	0	0	0	0	2	2	
40.	0	0	0	0	0	100	0	NA		1
41.	0	0	0	0	100	0	0	5	1	
42.	0	75	0	0	25	0	0	2	4	
43.	0	0	50	50	0	0	0	2	2	
44.	0	0	0	0	75	25	0	NA		2
45.	0	0	0	0	0	100	0	NA		1
46.	0	0	0	0	0	100	0	NA		1
47.	0	0	0	0	0	100	0	NA		1
48.	75	0	0	25	0	0	0	NA	4	
49.	25	25	0	0	50	0	0	2	5	
50.	0	25	0	0	75	0	0	3	4	
51.	0	0	0	0	100	0	0	5	1	
52.	0	0	0	0	0	100	0	NA		1
53.	0	0	25	25	50	0	0	4	3	
54.	0	0	0	0	75	25	0	NA		2
55.	0	0	0	0	0	100	0	4		1
56.	0	0	0	0	0	100	0	5		1
57.	0	0	0	0	0	100	0	4		1
58.	25	0	0	25	50	0	0	NA	5	
59.	0	50	0	25	25	0	0	3	4	
60.	0	25	0	0	75	0	0	3	4	
61.	0	25	0	0	75	0	0	5	4	
62.	0	0	0	0	0	100	0	NA		1
63.	0	50	50	0	0	0	0	2	2	
64.	0	0	0	0	75	25	0	3		2
65.	0	0	0	0	0	100	0	NA		1
66.	0	0	0	0	0	100	0	NA		1
67.	0	0	0	0	0	100	0	NA		1
68.	0	25	25	25	25	0	0	2	4	
69.	0	0	0	0	0	100	0	NA		1
70.	0	0	0	0	100	0	0	5	1	
71.	0	0	0	0	0	100	0	NA		1
72.	0	0	50	50	0	0	0	3	2	
73.	0	0	0	0	50	50	0	NA		2
74.	0	0	0	0	0	100	0	NA		1
75.	0	0	0	0	0	100	0	NA		1
76.	0	0	0	0	0	100	0	NA		1

The degree of agreement represented by the valuaes devoted to the last two columns in the table above "agreement among numerical responses" and "agreement among all reponses" as follows:

1 = Perfect agreement by all raters

2 = All ratings contained in 2 adjacent categories

3 = All ratings contained in 3 adjacent categories

4 = All ratings contained in 4 adjacent categories

5 = All ratings contained in 5 adjacent categories

Note: Because the two categorical reponses "NA" and "NR" could be scored differently in the agreement column if their order was reveresed in the table they are considered one category for the purpose of entering an agreement score in the last column.