ABSTRACT
              The guidance counselor is frequently seeking
information from the classroom teacher about the overt behavior of a
child in the classroom. In this study elementary school teachers were
asked to rate their students on items describing specific observable
classroom behaviors in two sessions with a two-week interval between
ratings. The items used to rate students were determined in a pilot
phase when elementary teachers were asked what concepts they
considered important and not important for the satisfactory behavior
of a child in the classroom. The results strongly indicated that
teachers were not stable in rating the overt behaviors of pupils. The
item reliabilities tended to increase as the number of rating
categories available to the teachers increased from five to seven to
nine; however, no statistical differences was found. Assuming that
teachers do rate and possibly refer children to an elementary
counselor based on a single episode, it would seem imperative that
the elementary counselor determine as quickly as possible the
generality of the behavior problem. (Author)

Paper Presented to a Meeting of the
American Personnel and
Guidance Association
March 1972

Stability of Teacher Ratings of Pupil Behaviors
in a Classroom Setting

Patricia B. Elmore and Donald L. Beggs
Southern Illinois University at Carbondale

STABILITY OF TEACHER RATINGS OF PUPIL BEHAVIORS

IN A CLASSROOM SETTING

Patricia B. Elmore and Donald L. Beggs

Southern Illinois University at Carbondale

The elementary counselor is responsible for obtaining

information from both students and teachers about activities

in the classroom setting. A great deal of work has been

done at the elementary school level in the process of

obtaining from pupils information that is accurate and

stable over a period of time. Fortunately, procedures

have been developed to assist us in data collection that

will give the counselor information from pupils concerning

their problems and the happenings in the classroom.

As a parallel to the investigation of the accuracy of

pupil reports of behaviors and attitudes, we have done very

little with respect to information that we are obtaining

from teachers concerning activities in the classroom.

There seems to be an unwritten law that the information

we obtain from teachers as it pertains to children's

behavior in the classroom is accurate and transcends over

the general classroom behavior of the child and does not

focus upon a single episode. There has been limited

research (Barnard, Zimbardo, & Sarason, 1968; Openshaw,

1967; Feshbach, 1969; and Tolor, Scarpetti, & Lane, 1967)

in the area of investigating the stability and the accuracy

of teachers' ratings especially as the rating relates to
pupil behaviors. In general, the results of the previous
research have indicated that the teachers are not consistent
in their ratings of pupil behavior.

The reasons for the lack of consistency in teacher
ratings of pupil behaviors have been discussed from several
viewpoints. Cronbach (1946, 1950) and Helmstadter (1957)
have suggested that "response style" has an undesirable
effect on the reliability of ratings. Although many
researchers (Conklin, 1923; Symonds, 1924; Champney & Marshall,
1939; Bendig & Hughes, 1953; Bendig, 1954; Garner, 1960; and
Eriksen & Hake, 1955a, 1955b) have investigated the optimal
number of rating categories, and there is no conclusive
evidence supporting any optimal number of scale categories.
Block (1957) and others have observed that rating scales
which do not encourage polarization or extreme responding
have, in general, very poor reliability.

The previous research has placed the practicing counselor
in a dilemma because the counselor is depending upon teacher
observation of classroom behavior of children. If the
teachers' ratings relate to a specific incident, then the
expectations of the counselor and what the counselor is to
do with the child are quite different than if the problem is
an acute problem that transcends all of the child's behavior.
Therefore, the counselor must be concerned with the type of
information he is obtaining from the classroom teacher. Is

the information stable with respect to a variety of situations

such as misbehavior in the classroom, or is the teacher

reporting information to the counselor based on a single

episode in the classroom? If the information is obtained

with respect to a specific episode, it would seem that

the counselor is highly restricted in the type of behavior

that might be expected in future performances of the child.

Therefore, this study has been undertaken to attempt to

determine if the information is accurate with respect to

children's behavior in the classroom as rated by the class-

room teacher.

Problem

This study was designed to investigate the manner in which

teachers respond to items measuring a concept judged

important and to items measuring a concept judged not

important by each teacher for five-, seven-, and nine-

category rating scales. A consideration in this study was

reliability of the items when employing five-, seven-, and

nine-category rating scales, i.e., the consistency of

teachers' responses to the same item on the three different

rating scales over repeated administrations.

Method

Subjects

The pilot sample consisted of twenty-nine elementary school

teachers working toward an advanced degree at Southern Illinois

University, Carbondale, Illinois. The sample for the major

study consisted of ninety-four teachers from Southern Illinois public

and parochial elementary schools.

Pilot Study

The first and second phases of the pilot study were
conducted to develop the appropriate instruments for the
major study. It was necessary to obtain concepts considered
important and concepts considered not important to elementary
school teachers for the satisfactory or acceptable behavior
of the child in the classroom. Also, the pilot study was
used to determine the concept that each of sixty items best
measured or described as perceived by elementary school
teachers. Sixteen concepts were included in the final
version of the Characteristics Scale.[1] The items, one
measuring each of the concepts, were combined in the same
order as the concepts they measured to form Behavior Rating
Scales I, II, and III.[2] These three rating scales were
constructed using the same items but with the number of
rating categories varied. Behavior Rating Scales I, II, and
III were five-, seven-, and nine-category rating scales
respectively. The rating categories were presented as a
continuum from Strongly Agree to Strongly Disagree with a
center category of No Comment.

Data Collection Procedure

For the major study each teacher was randomly assigned
to one of nine groups. Packets containing one copy of the
Characteristics Scale, the appropriate number of Behavior
Rating Scales I, II, or III, and the directions for the
scales were distributed to the teachers by the experimenter

during the scheduled meetings for the first testing period.
The teachers were verbally instructed by the experimenter
to open the packets of materials, read the written directions
for the scales, and then ask questions. The Directions for
the Characteristic Scale instructed each teacher to mark
the concepts in one of two categories, Important or Not
Important, according to his or her consideration of the
characteristic (concept) for the satisfactory behavior of
a child in the classroom. The Directions for Behavior
Rating Scales instructed the teacher to place an "X" in the
box along the continuum at the point which most nearly
described the student being rated with reference to the
behavior being considered. Each student in the teachers'
class was rated on the Behavior Rating Scale by the teacher.
For the first testing period, groups one, two, and three
received Behavior Rating Scale I; groups four, five, and
six received Behavior Rating Scale II; and groups seven,
eight, and nine received Behavior Rating Scale III. Two
weeks after the first set of materials was collected, an
appropriate number of one of the three Behavior Rating Scales
was distributed to each teacher individually by the exper-
imenter with the same written directions as the first testing
period. The teachers were not aware that they were going to
be asked to complete the materials this second time. There-
fore, there was no reason for the teachers to retain the
ratings they had given the first time. During the second

testing period groups one, four, and seven received Behavior

Rating Scale I; groups two, five, and eight received

Behavior Rating Scale II; and groups three, six, and nine

received Behavior Rating Scale III. Of the 94 teachers in

the original sample, 87 completed the study.

Results

The data obtained from the first and second testing

periods for teachers in groups one, five, and nine were

used to determine if teachers respond consistently over time

to each item on a five-, seven-, or nine-category rating

scale. The Pearson product-moment correlation coefficients

were computed to determine the item reliability for each

teacher in groups one, five, and nine for each of the

sixteen items on the five-, seven-, and nine-category

rating scales. These correlation coefficients were converted

using Fisher's logarithmic transformation of r to z_r values

which were averaged to obtain a z_r for each item on each of

the five-, seven-, and nine-category rating scales. The

obtained average z_r values were then converted to Pearson

product-moment correlation coefficients as shown in Table 1.

Insert Table 1 about here

The statistical hypotheses that population correlation

coefficients were not different from zero for each item on

each of the rating scales were retained at the .05 level of

significance with the exception of Item 1. The results

indicate, therefore, that teachers in this study did not

respond consistently over a short period of time (two

weeks between testing periods) when they rated each student

in their classrooms on each of sixteen items describing

specific behaviors related to a classroom setting.

The problem of determining the optimal number of

rating scale categories is conceptually related to the con-

sideration of stability of responding over time. In this

study a number of questions were generated concerning this

relationship. Is each item on a seven-category rating

scale more reliable over repeated administrations than the

corresponding item on a five-category rating scale? Is

each item on a nine-category rating scale more reliable

over repeated administrations than the corresponding item

on a seven-category rating scale? Is each item on a nine-

category rating scale more reliable over repeated administrations

than the corresponding item on a five-category rating scale?

The obtained z_r values were used to test the statistical

hypotheses that two populations have the same ρ-value.

The results are shown in terms of z values in Table 2.

Insert Table 2 about here

The hypotheses were retained at the .05 level of significance.

The results indicate that there is no statistically significant

difference in the item reliability of (a) each item on a

five-category rating scale and the corresponding item on a

seven-category rating scale, (b) each item on a seven-

category rating scale and the corresponding item on a

nine-category rating scale, and (c) each item on a five-

category rating scale and the corresponding item on a

nine-category rating scale. Although the statistical

analysis indicated that the number of categories on the

rating scale did not affect the reliability of the teachers'

ratings of their students on specific overt behaviors, a

visual observation of the reliability coefficients for each

item indicated that the reliability of responding increased

as the number of rating categories available to the teachers

increased. The data suggested this trend; however, the

trend was not analyzed statistically in this study.

Discussion

The results of this study clearly indicate that the

teachers' ratings of pupil behaviors over a short time

period are not stable. The study was developed such that

behaviors judged important by the elementary teachers were

included. Even the isolation of behaviors judged important

did not tend to stabilize the ratings.

The results indicate that the teachers are not rating

the general behaviors of the children on the rating sheets.

The teachers may very well be focusing upon a specific

episode involving a child when responding to a rating scale.

If this is true, a serious implication for elementary counselors

evolves. If referral rating sheets are completed based on a

single episode, the elementary counselor is being placed

in the role of a disciplinarian and not the role of a counselor.

The counselor cannot be expected to assist the child if the

counselor is fulfilling the role of a disciplinarian.

Assuming that teachers do rate and possibly refer

children to an elementary counselor based on a single

episode, it would seem imperative that the elementary counselor

determine as quickly as possible the generality of the

behavior problem. If the behavior problem is specific

to a single episode, the counselor should not have the

responsibility of dealing with the observed behavior. The

counselor should only be involved after the general nature

of the behavior problem has been established.

Elmore 10

References

Barnard, J. W., Zimbardo, P. G., & Sarason, S. B. Teachers'
ratings of student personality traits as they relate
to IQ and social desirability. Journal of Educational
Psychology, 1968, 59, 128-132.

Bendig, A. W. Transmitted information and the length of
rating scales. Journal of Experimental Psychology,
1954, 47, 303-308.

Bendig, A. W., & Hughes, J. B. Effect of amount of verbal
anchoring and number of rating-scale categories
upon transmitted information. Journal of Experi-
mental Psychology, 1953, 46, 87-90.

Block, J. Studies in the phenomenology of emotions. The
Journal of Abnormal and Social Psychology, 1957,
54, 358-363.

Champney, H., & Marshall, H. Optimal refinement of the
rating scale. Journal of Applied Psychology, 1939,
23, 323-331.

Conklin, E. S. The scale of values method for studies in
genetic psychology. University of Oregon Publica-
tion, 1923, 2, No. 1.

Cronbach, L. J. Response sets and test validity. Educa-
tional and Psychological Measurement, 1946, 6,
475-494.

Cronbach, L. J. Further evidence on response sets and
test design. Educational and Psychological
Measurement, 1950, 10, 3-31.

Eriksen, C. W., & Hake, H. W. Absolute judgments as a
          function of stimulus range and number of stimulus
          and response categories. Journal of Experimental
          Psychology, 1955, 49, 323-332. (a)

Eriksen, C. W., & Hake, H. W. Multidimensional stimulus
          differences and accuracy of discrimination.
          Journal of Experimental Psychology, 1955, 50,
          153-160. (b)

Feshbach, N. D. Student teacher preferences for elementary
          school pupils varying in personality character-
          istics. Journal of Educational Psychology, 1969,
          60, 126-132.

Garner, W. R. Rating scales, discriminability, and
          information transmission. Psychological Review,
          1960, 67, 343-352.

Helmstadter, G. C. Procedure for obtaining separate set
          and content components of a test score. Psycho-
          metrika, 1957, 22, 381-393.

Openshaw, K. A failure of the Minnesota Teacher Attitude
          Inventory to relate to teacher behavior. The
          Journal of Teacher Education, 1967, 18, 233-239.

Symonds, P. M. On the loss of reliability in ratings due
          to coarseness of the scale. Journal of Experi-
          mental Psychology, 1924, 7, 456-461.

Tolor, A., Scarpetti, W. L., & Lane, P. A. Teachers'
     .   attitudes toward children's behavior revisited.
          Journal of Educational Psychology, 1967, 58,
          175-180.

Footnotes

1. Copies are available upon request from the senior author.

2. Copies are available upon request from the senior author.

Table 1

Item Reliability Averaged Over Teachers In Groups One,

Five, And Nine For Each Of The Sixteen Items On The

Five-, Seven-, And Nine-Category Rating

Scales, Respectively

Item	Five-Category Rating Scale	Seven-Category Rating Scale	Nine-Category Rating Scale
1	.455	.600*	.575*
2	.020	.055	.095
3	-.035	.095	.200
4	.190	.265	.240
5	.045	.055	.180
6	.335	.345	.455
7	.070	.080	.215
8	-.020	.075	.190
9	.200	.145	.140
10	-.060	.100	-.050
11	-.045	.095	.195
12	.025	.080	-.035
13	.060	.235	.155
14	.085	.160	.045
15	.115	.190	.280
16	-.050	.225	.140

*A sample correlation value of .549 was required

for the statistic to be significant at the .05 level of

significance.   14

Table 2

z-Values Obtained From Testing The Hypotheses That

Two Rating Scales With Different Number Of

Categories Have The Same Item Reliability

Item	Five Categories vs. Seven Categories	Seven Categories vs. Nine Categories	Five Categories vs. Nine Categories
1	-.38	+.07	-.30
2	-.07	-.07	-.14
3	-.25	-.20	-.44
4	-.15	+.05	-.10
5	-.02	-.24	-.26
6	-.02	-.25	-.27
7	-.02	-.26	-.28
8	-.19	-.21	-.41
9	+.10	+.01	+.11
10	-.30	+.29	-.02
11	-.27	-.19	-.46
12	-.10	+.22	+.12
13	-.34	+.16	-.18
14	-.14	+.21	+.07
15	-.15	-.17	-.32
16	-.52	+.17	-.36

DIRECTIONS FOR CHARACTERISTICS SCALE

1. Print your full name, the name of the elementary school in which you are teaching, and the city in which the school is located on the attached sheet. Be sure to indicate the grade you are presently teaching.

2. Please read the characteristics carefully. This list of sixteen characteristics was determined by a group of elementary school teachers who considered some of the characteristics to be important and others to be not important.

3. Indicate by a check mark ($\checkmark$) the characteristics, listed on the attached sheet, that you consider important and not important considerations for the satisfactory (or acceptable) behavior of a child in the classroom.

4. In order to determine whether a characteristic is important or not important to you, think in terms of the characteristics of pupils you enjoy teaching. The word "behavior" does NOT refer to academic success; it refers to how the child ACTS in the classroom.

5. There is no time limit.

16

CHARACTERISTICS SCALE

Name _____

School _____

City _____

Grade You Are Teaching _____

IMPORTANT	NOT IMPORTANT	CHARACTERISTICS
____	____	1. Aggressive, tends toward fighting, bullying, teasing, cruelty, vs. non-aggressive, kind, considerate.
____	____	2. Demanding of teacher's attention, vs. prefers not to be noticed.
____	____	3. Of generally good health, vs. poor general health.
____	____	4. Irresponsible, frivolous, vs. responsible.
____	____	5. Self-assertive, tends to dominate other children, vs. submissive, follows lead of other children.
____	____	6. Popular, generally liked by other children, vs. unpopular, generally disliked by other children.
____	____	7. Cooperative, compliant, courteous with children and adults, vs. negativistic, stubborn, disobedient, discourteous, argumentative, "poor sport".
____	____	8. Good posture, vs. poor posture.
____	____	9. Self-centered, conceited, boastful, "show-off", vs. self-abasive, deferent, minimizes own importance.
____	____	10. Associates primarily with children of own sex, vs. associates primarily with children of opposite sex.
____	____	11. Physically strong, vs. physically weak.
____	____	12. Stable in interests, attitudes, opinions, vs. changeable.
____	____	13. Gregarious, prefers games involving many children, vs. prefers solitary pursuits.
____	____	14. Quiet, vs. talkative, distracting in class.
____	____	15. Good tonal quality of voice, vs. bad tonal quality of voice.
____	____	16. Learns fast, vs. learns slowly.

DIRECTIONS FOR BEHAVIOR RATING SCALES

This inventory consists of sixteen statements designed to sample your opinions about your pupils and their behavior in the classroom. There are no right or wrong answers except that they are your own opinions. What is wanted is your own individual feeling about each student for the statements. Read each statement and decide how YOU feel about it.

Place an X in the box, ☒ , along the continuum Strongly Agree to Strongly Disagree at the point which most nearly describes the student with reference to the BEHAVIOR you are considering. PLEASE RESPOND TO EVERY ITEM.

BEHAVIOR RATING SCALE I

Student's Name _____

1. _____ teases other pupils.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

2. _____ demands the teacher's attention.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

3. _____ has poor general health.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

4. _____ is irresponsible.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

5. _____ dominates other children.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

6. _____ is unpopular.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

7. _____ is disobedient.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

8. _____ has poor posture.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

9. _____ is a show-off.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

10. _____ associates primarily with children of opposite sex.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

11. _____ is physically weak.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

12. _____ is changeable.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

13. _____ prefers games involving many children.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

14. _____ is distracting in class.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

15. _____'s voice has bad tonal quality.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

16. _____ learns slowly.

☐ Strongly Agree ☐ ☐ No Comment ☐ ☐ Strongly Disagree

19

BEHAVIOR RATING SCALE II

Student's Name _____

1. _____ teaaes other pupils.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

2. _____ demands the teacher's attention.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

3. _____ has poor general health.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

4. _____ is irresponsible.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

5. _____ dominates other children.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

6. _____ is unpopular.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

7. _____ is disobedient.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

8. _____ has poor posture.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

9. _____ is a show-off.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

10. _____ associates primarily with children of opposite sex.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

11. _____ is physically weak.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

12. _____ is changeable.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

13. _____ prefers games involving many children.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

14. _____ is distracting in class.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

15. _____'s voice has bad tonal quality.

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

16. _____ learns slowly.   20

☐ ☐ ☐ ☐ ☐ ☐ ☐
Strongly Agree   No Comment   Strongly Disagree

BEHAVIOR RATING SCALE III

Student's Name _____

1. _____ teases other pupils.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

2. _____ demands the teacher's attention.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

3. _____ has poor general health.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

4. _____ is irresponsible.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

5. _____ dominates other children.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

6. _____ is unpopular.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

7. _____ is disobedient.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

8. _____ has poor posture.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

9. _____ is a show-off.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

10. _____ associates primarily with children of opposite sex.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

11. _____ is physically weak.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

12. _____ is changeable.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

13. _____ prefers games involving many children.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

14. _____ is distracting in class.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

15. _____'s voice has bad tonal quality.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ ☐ Strongly Disagree

16. _____ learns slowly.

☐ Strongly Agree ☐ ☐ ☐ ☐ No Comment ☐ ☐ ☐ 21 ☐ ☐ Strongly Disagree