1981

Reliability and validity of a scale to measure prosocial behavior in young children

Susan Davis
Portland State University

AN ABSTRACT OF THE THESIS OF Susan Davis for the Master of Science in Psychology presented November 9, 1981.

Title: Reliability and Validity of a Scale to Measure Prosocial Behavior in Young Children.

APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

[redacted]

Cathleen L. Smith, Chairperson

[redacted]

Barbara J. Stewart

[redacted]

Hugo M. Maynard

The present study was designed to determine the reliability and validity of an observation code and rating scale developed by Smith (unpublished research) to measure prosocial behavior in young children.

Forty-two individual children ($\bar{x}$ age=53 months) were given opportunities to behave prosocially (i.e., teach, help, share, comfort) during a naturalistic play session with two adult experimenters. Two hidden observers (referred to as trained raters) observed and rated the child's prosocial responses using the observation code and rating scale developed by Smith (unpublished research) according to the following general format: 1--no response, 2--concern with no involvement, 3--concern which poses a

solution to the need, 4--prosocial responses (i.e., teaching, helping, sharing, comforting), 5--prosocial responses with special involvement. Videotapes were made of each child's play session and prosocial responses. These videotapes were then rearranged so that all like behaviors were clustered together. For example, all helping responses were assembled on one tape, all teaching responses on another tape, and so on. These tapes are referred to as the main tapes. Fifty-five undergraduate psychology students, acting as untrained raters, viewed and rated the teaching and helping tapes (25 of the untrained raters) or the sharing and comforting tapes (30 of the untrained raters). The untrained raters were asked to rate each child's response on a 5-point scale, from lowest to highest amount of prosocial behavior. In addition, a short tape containing the prosocial responses of 12 to 15 children was constructed for each behavioral category (i.e., teaching, helping, sharing, comforting). These sample tapes were shown just prior to the main tape to give the subjests an opportunity to see and rate a sample of the range of responses for that behavior; the same short tape (referred to as the explanation tape) was presented again following the main tape to provide the subjects with a second opportunity to assign a rating and provide an explanation for their choice of that rating for each child.

Results indicated that when given minimal guidelines untrained raters showed a significant amount of agreement among themselves when rating the prosocial responses of young children, particularly on the sharing, teaching, and helping tapes. Agreement on comforting, although still significant, was somewhat lower. In addition, and most importantly, the ratings of untrained raters were highly correlated for all behaviors with the ratings of persons skilled in the use of systematic observational

methods. Further, when the explanations given by the untrained raters were subjected to a content analysis, they were found to be highly comparable to the guidelines contained in the observation code used by the trained raters. Therefore, it appears that the observation code and rating scale developed by Smith (unpublished research) is a reliable and valid measure of prosocial behavior in young children.

RELIABILITY AND VALIDITY OF A SCALE TO MEASURE PROSOCIAL BEHAVIOR IN YOUNG CHILDREN

by

SUSAN DAVIS

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

PSYCHOLOGY

Portland State University

1981

TO THE OFFICE OF GRADUATE STUDIES AND RESEARCH:

The members of the Committee approve the thesis of Susan Davis presented November 9, 1981.

[redacted]

Cathleen L. Smith, Chairperson

[redacted]

Barbara J. Stewart

[redacted]

Hugo M. Maynard

APPROVED:

[redacted]

David F. Wrench, Head, Department of Psychology

[redacted]

Stanley E. Rauch, Dean of Graduate Studies and Research

ACKNOWLEDGMENTS

My sincere thanks go to Cathleen Smith, without whose guidance, support, and empathy this project could not have been done. I also wish to thank Barbara Stewart for the many hours she donated unselfishly to this study and to my own statistical knowledge. In addition, I wish to thank Hugo Maynard for his dedicated and enlightening contributions to this project.

Karen King deserves a special thank you for her kindness and time utilized in the development of the computer programming.

I also want to thank my son J.P. Elliott, whose existence made me want to be someone he could be proud of. And finally, although words are not enough, my thanks to Dick for the many years he has supported, encouraged, and had faith in me.

TABLE OF CONTENTS

		PAGE
ACKNOWLEDGMENTS	. . . . .	iii
LIST OF TABLES	. . . . .	v
LIST OF FIGURES	. . . . .	vi
CHAPTER		
I	INTRODUCTION AND LITERATURE REVIEW	1
II	METHOD	9
	Overview	9
	Subjects	9
	Apparatus	9
	Rating Forms	11
	Procedure	12
III	RESULTS	15
	Interrater Reliability of Untrained Raters. .	15
	Interrater Reliability of Trained Raters . . .	20
	Comparison: Trained and Untrained Raters . .	20
	Content Analysis of Untrained Raters' Explanations, and Comparison with Observation Code Used by Trained Raters	29
IV	DISCUSSION	35
REFERENCE NOTES	. . . . .	41
REFERENCE LIST	. . . . .	42
APPENDIX	. . . . .	43

LIST OF TABLES

TABLE		PAGE
I	Median Correlations of the Untrained Raters - Teaching. .	16
II	Median Correlations of the Untrained Raters - Helping . .	17
III	Median Correlations of the Untrained Raters - Sharing . .	18
IV	Median Correlations of the Untrained Raters - Comforting .	19
V	Rater Reliabilities of Two Trained Raters for Teaching, Helping, Sharing, and Comforting	21
VI	Pearson Correlations Between Average Trained and Average Untrained Raters' Ratings for Teaching, Helping, Sharing, and Comforting	23
VII	Means, Medians, Standard Deviations, and T-tests for Trained and Untrained Raters on the Main Tapes for All Behaviors	24
VIII	Comparison of Untrained Raters' Explanations with Observation Code Used by Trained Raters: Teaching . .	30
IX	Comparison of Untrained Raters' Explanations with Observation Code Used by Trained Raters: Helping . . .	31
X	Comparison of Untrained Raters' Explanations with Observation Code Used by Trained Raters: Sharing . . .	32
XI	Comparison of Untrained Raters' Explanations with Observation Code Used by Trained Raters: Comforting .	33

LIST OF FIGURES

FIGURE PAGE

1. Distribution of Average Ratings of Trained and Untrained Raters for Teaching Main Tape 25

2. Distribution of Average Ratings of Trained and Untrained Raters for Helping Main Tape 26

3. Distribution of Average Ratings of Trained and Untrained Raters for Sharing Main Tapes 27

4. Distribution of Average Ratings of Trained and Untrained Raters for Comforting Main Tapes 28

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Since the early 1960's developmental researchers have been interested in children's inclinations to intervene on behalf of another. This intervention has been termed prosocial behavior. Prosocial behavior can be defined as a child's willingness and ability to come to the aid of another, often at some cost to self. Although this behavior has typically been measured in contrived laboratory situations using a single dependent measure, such as the number of marbles or gift certificates donated to an absent needy peer, a few investigators have attempted to measure prosocial behavior in more naturalistic situations.

One such investigation was conducted by Yarrow, Scott and Waxler (1973). In a test designed to measure whether symbolic prosocial training using pictures and dioramas generalized to real-life prosocial opportunities, children were exposed to two natural-appearing situations in which they could come to the aid of an adult and an infant. Specifically, each child was invited to come to a house next door to the site where the symbolic learning had taken place. There a mother and a year-old baby were visiting. While waiting for the baby to be ready, an assistant and the child sat down to look at some picture books. While they were looking at the books, a basket of spools and buttons "accidentally" fell off the table. The dependent measure was whether the child helped pick up the spools. After sufficient time had elapsed for the spools to be gathered by the child, the mother called to the child to come over to the playpen, and together they played with the infant. Then the mother asked the child to look after the

infant while she went to get juice and crackers. As she left, she picked up a blanket from the floor and exposed some toys which had fallen out of the playpen, thus presenting the child with a second opportunity to help, this time by picking up the toys and giving them to the baby. A snack followed and the visit was pleasantly terminated. Records of the child's responses to these staged prosocial opportunities were made by an adult seated unobtrusively at a desk in a far corner of the room.

In a further study, Yarrow and Waxler (1976) refined their measurement system by adding an observation code for rating prosocial behavior in young children. Children were exposed in a naturalistic play setting to a number of situations in which they could intervene on behalf of an adult who experienced "distress". For instance, in one session it became apparent that there were not enough Cheerios for the adult experimenter. She commented about the inequality and expressed disappointment in her small serving, thereby presenting the child with an opportunity to share. In a helping opportunity the experimenter "accidently" spilled a box of tennis balls. She made no comment but appeared occupied with other materials, thus giving the child an opportunity to pick up the fallen balls. An opportunity for comforting occurred when the experimenter "accidently" pinched her finger in a drawer. She grimaced and held her finger. Children's responses to these prosocial opportunities were recorded by observers on a 6-point scale: 1--active indifference, 2--no apparent notice, 3--recognition only, 4--concern and/or partial attempts to intervene, 5--aid, and 6--aid with special involvement.

In a study using a method adapted from Yarrow, Scott and Waxler (1973), and Yarrow and Waxler (1976), Smith, (Reference Note 1) provided a small number of children with opportunities to help, share with, or comfort an

adult experimenter. In addition, teaching, which had not been examined by Yarrow and her colleagues, was included by Smith in response to Staub's (1975) suggestion that this behavior, regardless of content, is a prosocial response. Opportunities for the child to teach occurred when one of the experimenters indicated she did not know how to do something (e.g., "I don't know how to make colored playdough"). In the teaching opportunities the child was always asked to demonstrate a simple skill or impart information which he or she had just learned from the other experimenter. For all prosocial behaviors, observers trained in systematic observation recorded each child's response as a running narrative.

In later pilot research (Smith, unpublished research) the responses of a large number of children to these staged, yet natural-appearing prosocial opportunities were examined, and the narrative format was modified to an objective 5-point rating scale using the following general categories: 1--the child displays no physical or verbal response to the adult's need for help, comfort, sharing or teaching, 2--the child acknowledges the adult's need but does not actually engage in behavior to correct the situation, 3--the child indicates verbally that some specific action would improve or correct the situation but does not engage in the behavior herself/himself, 4--(help) the child helps in a task which is better done by more than one person; (share) an object belonging to the child is shared with the adult to correct an inequality; (teach) the child teaches the naive adult a three-part skill; (comfort) the child expresses sympathy, physically or verbally, concerning the adult's injury, 5--prosocial responses with special involvment: (help) the child helps immediately and completely; (share) the child shares all remaining articles in his/her possession; (teach) the child demonstrates

physically and verbally all components of the teaching skill; (comfort) the child verbalizes a sympathetic response and physically demonstrates concern (i.e., a kiss or pat), or verbalizes extensive sympathetic responses. The complete definitions of each observational scoring category are presented in the Appendix.

In 1980, Blackwell, Smith and Stewart (Reference Note 2) observed preschool children's prosocial behavior (i.e., sharing, teaching, helping, and comforting) in naturalistic play situations. As before, children were offered standardized opportunities to behave prosocially in an experimental context which was designed to be maximally similar to situations actually encountered in their daily lives. Individual children were escorted by two female experimenters to a familiar playroom in their school where they engaged in playful activities (e.g., planting seeds, playing with playdough) into which a number of everyday situations involving the opportunity to help, teach, share, and comfort were blended. Extensive pretesting enabled the experimenters to devise ways of presenting the prosocial opportunities as naturally and unobtrusively as possible by taking advantage of materials and activities at hand. Opportunities to behave prosocially were introduced by means of statements, termed "verbalizations of need," which were designed to provide relatively unambiguious indications that the experimenter was in need; in all cases, however, the need verbalizations stopped short of directly asking the child to intervene. For example, helping opportunities were provided when one of the experimenters appeared to have accidently dropped objects (e.g., "Oh! I spilled the sticks"), when objects ostensibly became lost (e.g., "My special box is lost"), or when materials needed to be moved or cleaned up. Sharing opportunities

occurred when the experimenter indicated her desire for an object in the child's possession that had previously been labeled as belonging to the child (e.g., "I'd like to plant seeds, but I don't have a cup"). Teaching opportunities occurred when one of the experimenters acknowledged that she did not know how to do a simple three-part task (e.g., "I don't know how to make a flower stick") which the child had just been taught by the other experimenter (e.g., to make a flower stick, one must lick the back of the flower sticker, press it firmly at one end of a "popsicle" stick, and color the stick green with a green felt marker). Finally, an opportunity to comfort was provided when one of the experimenters appeared to injure herself accidently (e.g., "Oh! I bumped my knee, It really hurts") and demonstrated mild distress and appropriate nonverbal cues.

In this study four observers located behind a cabinet with a one-way mirror recorded and rated the level of the child's responses using the observation code and 5-point scale developed by Smith. Inter-observer reliability for each of the prosocial behaviors was calculated using percent agreement. Exact agreement of the observers was 86% for helping, 86% for sharing, 82% for teaching, and 76% for comforting. By including those disagreements within one point on the 5-point scale, inter-rater reliability reached 98% for helping, 98% for sharing, 98% for teaching, and 98% for comforting.

It is clear from these reliability figures that the 5-point scale and observation code developed by Smith was used in approximately the same way by the four observers in the Blackwell et al. (Reference Note 2) study. However, these four observers had had extensive experience with the responses of young children in these situations and abundant training in systematic observational methods. Would people with less training and experience

view the children's behavior in these standardized situations as prosocial? When given minimal guidelines, would people in general show agreement among themselves when rating young children's responses in these situations? Would the ratings of people in general agree with the ratings of observers trained in systematic observational methods and the use of this observation code and rating scale? And finally, would people in general express the same reasons for assigning a rating at any given scale point as were contained in the observation code used by the trained raters? These questions address themselves to the issues of validity and reliability.

Validity refers to the ability of an instrument to measure what it was designed to measure. However, Nunnally (1967) states that the purpose of validity is not to assess the instrument (in this case the observation code) but the use to which the instrument is put (i.e., measurement of the level of children's responses to prosocial opportunities). Reliability refers to the consistency of measurement throughout a series of similar instruments (Cronbach, 1960). That is, in order to assess the reliability of an instrument results must be obtained from a series of similar instruments. These measurements must then correlate (agree) to a statistically significant degree.

The issues of reliability and validity have important implications for further use of the observation code and rating scale developed by Smith and her collegues. While much research has examined the prosocial responses and behavior of young children, only rarely have researchers rated the children's responses on a continuum from least to most prosocial. Instead, a majority of researchers use a single dependent measure in which the rating of the children's behavior is entirely constrained (e.g., the child donated to a needy peer or not, or the child picked up the dropped objects or not).

It therefore seems important to create a standardized observation code and rating scale which is not only accurate but is designed to measure a range of prosocial behavior in young children. The observation code developed by Smith (unpublished research) utilizes a 5-point rating scale which allows an observer to rate the level or amount of prosocial behavior a child shows. Once such an observation code is developed it is imperative that it's reliability and validity be tested. In addition, it is necessary to determine the ease and accuracy with which this observation code could be used. Once these factors have been determined, the results of studies of children's prosocial behavior by different researchers could be compared in a more comprehensive manner.

The present study was designed to determine the reliability and validity of the observation code and rating scale developed by Smith to measure prosocial behavior in young children, and to meet the need of a standardized observation code to be used by other researchers studying prosocial behavior in young children. A brief description of this study is outlined below.

Forty-two individual children ($\bar{x}$ age=53months) were given opportunities to behave prosocially using the Blackwell et al. (Reference Note 2) method. Two hidden observers (hereinafter referred to as the trained raters) independently rated the child's prosocial responses using the observation code and 5-point rating scale described earlier. Videotapes were made of each child's prosocial behavior. The videotapes were then rearranged so that all like behaviors were clustered together. For example, the helping opportunities for all children were assembled on one tape, all teaching opportunities on another tape, and so one. The videotapes were then shown to 55 undergraduate students (hereinafter referred

to as the untrained raters) who were asked to rate each child's response for amount of prosocial behavior on a scale from 1 to 5, 1 being the lowest level of response and 5 being the highest level of response. Interrater reliability was computed for the untrained raters, and their ratings were then compared to those obtained from the trained raters. In addition, differences in scoring between the trained and untrained raters were examined qualitatively.

CHAPTER II

METHOD

Overview

Fifty-five undergraduate psychology students at Portland State University, acting as untrained raters, viewed and rated for amount of prosocial behavior the videotaped responses of individual children to opportunities to teach and help (25 of the untrained raters) or share and comfort (30 of the untrained raters). Students (untrained raters) were asked to rate each child's response on a 5-point scale, from lowest to highest amount of prosocial behavior. The ratings of the untrained raters were then compared to the ratings of the same prosocial responses by raters trained in systematic obseration.

Subjects

Subjects were 55 students at Portland State University recruited from undergraduate psychology courses. Participants were given extra credit in the class in which they were enrolled, as per Psychology Department policy.

Apparatus

A series of videotaped incidents was compiled showing the responses of 42 children to opportunities to share, teach, help, and comfort, as defined in Chapter I. Each of the 42 original videotapes was a record of the child's responses to an experimental session in which several prosocial opportunities were embedded as naturally as possible in an ongoing play

activity, planting seeds. Permission to use these videotapes for research purposes had been previously granted by the children's parents. The videotaped experimental session presented the child with three opportunities each for sharing, teaching, and helping, and one opportunity for comforting. Of the 420 total (i.e., 42 children were given three opportunities to help, teach, share, and one to comfort, a total of ten opportunities per child) videotaped prosocial responses, 149 were selected for use in this study. A 1,2,3, random method was employed to select the video segments to be used. For example, for child 001 the first helping opportunity, the first sharing opportunity, and the first teaching opportunity was used; for child 002 helping 2, sharing 2, and teaching 2 was used; for child 003 helping 3, sharing 3, and teaching 3 was used, and so on. Since there was only one opportunity in the session for the child to comfort, children's comforting responses were selected randomly using the random numbers table. All like behaviors were then clustered together to construct new tapes. For example, helping responses of all 42 children appear on one tape, the 42 teaching responses on a second tape, and the 42 sharing responses on a third tape. A final tape consisting of 20 rather than 42 comforting responses was also constructed. Because so many of the 42 children exhibited no comforting responses, 22 of these children were eliminated from the comforting tape in order to prevent boredom among the untrained raters. Together these four tapes are hereinafter referred to as the <u>main</u> tapes.

Once constructed, each main tape was checked to insure that it contained the same proportion of prosocial responses per scale point as was contained in the original sample of 420 prosocial respones. In other words, if, of the original 129 helping responses one-fifth were scored by the trained raters as a scale point 2 response, then the newly constructed main helping tape was inspected to insure that one-fifth of the responses had also been rated at

scale point 2. In each of the main tapes for helping, teaching, sharing and comforting the proportions of ratings remained the same as in the original sample. However, due to a low proportion of scale point 3 responses in the original sharing sample, an additional scale point 3 response was randomly selected for the main sharing tape in order to obtain more information about how untrained raters would rate children's sharing responses at that scale point.

In addition, a short tape containing the prosocial responses of 12 to 15 children was constucted for each behavioral category (i.e., teaching, helping, sharing, comforting). For each behavioral category, the corresponding short tape (referred to as the sample tape) was shown just prior to the main tape to give the subjects an opportunity to see and rate a sample of the range for that behavior; the same short tape (referred to as the explanation tape) was presented again after the main tape to provide the subjects with a second opportunity to assign a rating and provide an explanation for their choice of that rating for each child. It will be remembered that these behavior sequences had been previously rated in vivo by trained raters. Using the ratings for each child's response obtained from the trained raters, the sample tape for each behavior included at least two examples of each scale point on the 5-point rating scale. For example, the sample helping tape contained two helping responses rated by the trained raters as a scale point 1 response, two helping responses rated as a 2, and so on for each scale point.

Rating Forms

For each tape, subjects were given scoring sheets appropriately numbered, with instructions to circle the chosen rating from 1 to 5 for each child's prosocial response. Additionally, for the explanation tapes, subjects were asked to circle a rating for each child and describe in the space provided what it

was about that child's behavior that made them choose that rating.

Procedure

To assure that the time requirement for each subject was not prohibitive, 25 subjects viewed the teaching and helping tapes, while the remaining 30 subjects viewed the sharing and comforting tapes. The tapes were shown to subjects in groups of five to seven students. For each group of subjects on each behavior, the sample tape was shown first, followed by the main tape, and finally the explanation tape.

Subjects were seated at a table facing the videotape monitor. Dividers were placed between each subject to insure independent ratings. The experimenter told subjects that they were to view videotapes of the responses of 42 children to opportunities to help and teach or share and comfort. Subjects were instructed that they were to rate each child's response (on the scoring sheets provided preceeding each tape) for amount of prosocial behavior shown on a 5-point scale, with 1 being the lowest level of response and 5 being the highest level of response. The experimenter informed the subjects that a 12-second delay between each child's response had been provided for rating purposes.

The experimenter then cautioned the subjects as to the importance of protecting the anonymity of the children viewed in the tapes. The investigator stressed that in the event a subject, while functioning as a rater, recognized a child in the tapes, it would be extremely important to protect that child's privacy by not repeating anything seen or heard in the videotapes.

The experimenter asked the subjects to use their own judgments in rating, not to be concerned with how their fellow subjects were rating, and finally that their ratings were individual judgments with no right or wrong answers. Subjects were told that to assist in their ratings the adult exper-

imenter, shown in the tapes, would repeat any verbalizations made by the child which were unclear or spoken too softly to be easily heard.

The investigator gave the subjects a short explanation of the events preceding the videotaped responses they were to rate. For example, subjects were told that prior to the videotaped responses for sharing the adult experimenter (shown with the child in the tape) had indicated a desire for an object or material in the child's possession and previously labeled as belonging to the child (e.g., "I'd like to plant seeds, but I don't have a cup"; "I'd like to make a flower stick, but I don't have a flower"; "I really like animal crackers, but I don't have any"). In the requests for teaching, the adult experimenter had acknowledged that she did not know how to do something (e.g., "I don't know how to make a flower stick"). Subjects were told that each teaching opportunity pertained to a simple three-part skill or task the child had learned from the other adult experimenter, and the three components for each teaching opportunity were then described.

The complete video segments for helping and comforting were shown the raters, so that no previous explanation was necessary. For instance, the videotaped segment for helping began when the adult experimenter said, "Oh, the bags on the table need to be moved to the suitcase". It was evident in the tapes that the adult experimenter had her arms full of supplies and was unable to pick up the bags herself. Comforting opportunities began when the adult experimenter bumped her knee when sitting down and said, "Oh, I bumped my knee, it really hurts". The videotaped segments, seen by the raters, began prior to the adult bumping her knee.

To assist the subjects in rating each child's prosocial response, some general guidelines for the rating procedure were outlined. For instance, the subjects were instructed to observe movement of the child, physical contact with the adult, eye contact with the adult, expressions on

the child's face and any verbalizations by the child. Subjects were instructed to use everything they had observed to obtain an overall or global impression of the child's level of teaching, helping, sharing or comforting, and to rate accordingly on the 5-point rating scale.

To illustrate a typical experimental session, one group of five to seven subjects viewed and rated the sample, main, and explanation tapes for sharing. After a short break the same group of subjects viewed and rated the sample, main, and explanation tapes for comforting. Other groups followed the same procedure for the helping and teaching tapes. The presentation of the tapes was alternated so that half of the groups viewed either the comforting or the helping sequences first, while the remaining groups viewed either the sharing or the teaching tape sequences first. At the end of each session each group of subjects was again cautioned to maintain the confidentiality of the children in the tapes, and thanked for their time and participation in the study.

CHAPTER III

RESULTS

Interrater Reliability of Untrained Raters

Interrater reliabilities for each untrained rater were obtained by correlating his or her ratings of responses on each tape (sample, main, explanation) for a given behavior (helping, teaching, sharing, comforting) with the ratings of every other untrained rater viewing that behavior tape. For example, for each of the teaching tapes (sample, main, explanation), the ratings of each untrained rater were correlated with the ratings of each of the remaining 24 raters, resulting in 24 correlations for each rater for each tape. To indicate how well each rater agreed with the other 24 raters on each of the three teaching tapes, the median of each set of 24 correlations was determined. In Table I then, the three entries for each untrained rater are his or her three median rater reliabilities for the teaching sample, main, and explanation tapes. This procedure was repeated to obtain the three median rater reliabilities for each of the 25 untrained raters on the helping sample, main, and explanation tapes (See Table II). The results in Tables III and IV represent the median rater reliabilities of the 30 untrained raters who viewed the sharing sample, main, and explanation tapes and the comforting sample, main, and explanation tapes.

When the ratings of each untrained rater were correlated with the ratings of each of the other untrained raters, agreement was highest for teaching, helping, and sharing across all tapes (i.e., sample, main, explanation). Agreement among raters was somewhat lower on comforting, particularily on the main tape.

TABLE I

MEDIAN CORRELATIONS OF THE UNTRAINED RATERS

TEACHING

UNTRAINED RATER	TEACHING SAMPLE TAPE[a]	TEACHING MAIN TAPE[b]	TEACHING EXPLANATION TAPE[c]
1	.8450	.7588	.8744
2	.8005	.8143	.9048
3	.8177	.7682	.8632
4	.7119	.7757	.8330
5	.7629	.7382	.7731
6	.7175	.8083	.8649
7	.7762	.7513	.8032
8	.8240	.8483	.8402
9	.7227	.8274	.8728
10	.8226	.8395	.7474
11	.8230	.8250	.8802
12	.8617	.8605	.8869
13	.8629	.8375	.9123
14	.8406	.8492	.8570
15	.8300	.8777	.8188
16	.8240	.8552	.8676
17	.7574	.8164	.8960
18	.8076	.8141	.8459
19	.7066	.7329	.8662
20	.7468	.7016	.8314
21	.8384	.8198	.9187
22	.6873	.7528	.7498
23	.8417	.8437	.8744
24	.7926	.8337	.8092
25	.7578	.7856	.8765

Note. For each of the tapes (sample, main and explanation) the ratings of each untrained rater were correlated with the ratings of each of the remaining 24 raters, resulting in 24 correlations for each rater for each tape. To indicate how well each rater agreed with the other 24 raters, on each of the three tapes the median correlation was determined. Entries in the above Table are these median rater reliabilities.

[a]For the teaching sample tape, 14 behavioral responses were rated. Any $r = .5324$ is significant at $p < .025$; $r = .6614$ is significant at $p < .005$; $r = .7800$ is significant at $p < .0005$.

[b]For the teaching main tape, 42 behavioral responses were rated. Any $r = .3044$ is significant at $p < .025$; $r = .3932$ is significant at $p < .005$; $r = .4896$ is significant at $p < .0005$.

[c]For the teaching explanation tape, 14 behavioral responses were rated. Any $r = .5324$ is significant at $p < .025$; $r = .6614$ is significant at $p < .005$; $r = .7800$ is significant at $p < .0005$.

TABLE II

MEDIAN CORRELATIONS OF THE UNTRAINED RATERS

HELPING

UNTRAINED RATER	HELPING SAMPLE TAPE	HELPING MAIN TAPE	HELPING EXPLANATION TAPE
1	.7670	.8763	.8877
2	.8566	.8592	.7482
3	.8947	.8875	.9097
4	.8740	.8697	.8129
5	.7880	.7679	.8588
6	.9041	.8603	.8646
7	.9169	.8631	.8408
8	.8947	.8276	.7622
9	.8953	.8982	.7950
10	.8081	.8154	.8977
11	.9008	.8617	.8960
12	.8776	.8800	.9122
13	.7989	.8408	.8919
14	.7549	.9026	.8954
15	.7990	.8689	.8999
16	.8686	.8490	.8648
17	.9100	.8858	.9155
18	.8219	.8380	.8129
19	.8509	.8490	.8841
20	.6604	.8199	.8070
21	.8214	.8528	.8877
22	.8665	.7941	.8123
23	.8512	.8590	.8877
24	.8522	.7806	.4129
25	.8803	.8697	.8741

Note. For each of the tapes (sample, main and explanation) the ratings of each untrained rater were correlated with the ratings of each of the remaining 24 raters, resulting in 24 correlations for each rater for each tape. To indicate how well each rater agreed with the other 24 raters, on each of the three tapes the median correlation was determined. Entries in the above Table are these median rater reliabilities.

[a]For the helping sample tape, 15 behavioral responses were rated. Any $r = .5139$ is significant at $p < .025$; $r = .6411$ is significant at $p < .005$; $r = .7603$ is significant at $p < .0005$.

[b]For the helping main tape, 42 behavioral responses were rated. Any $r = .3044$ is significant at $p < .025$; $r = .3932$ is significant at $p < .005$; $r = .4896$ is significant at $p < .0005$.

[c]For the helping explanation tape, 15 behavioral responses were rated. Any $r = .5139$ is significant at $p < .025$; $r = .6411$ is significant at $p < .005$; $r = .7603$ is significant at $p < .0005$.

TABLE III

MEDIAN CORRELATIONS OF UNTRAINED RATERS SHARING

UNTRAINED RATER	SHARING SAMPLE TAPE	SHARING MAIN TAPE	SHARING EXPLANATION TAPE
1	.9150	.9573	.9356
2	.8448	.9497	.8681
3	.9348	.8022	.9340
4	.9093	.9684	.9104
5	.8878	.7423	.8231
6	.9026	.9647	.8546
7	.8929	.9452	.8698
8	.8665	.9417	.8515
9	.9199	.9613	.9366
10	.8281	.9409	.9205
11	.8999	.9540	.9435
12	.9192	.9666	.9259
13	.7378	.9340	.8877
14	.8383	.9430	.8850
15	.9075	.9288	.8850
16	.9146	.9507	.5961
17	.9267	.9655	.9276
18	.4611	.9273	.9286
19	.9265	.9497	.9340
20	.9006	.9310	.9105
21	.9281	.9461	.9313
22	.9281	.9461	.9313
23	.9213	.9591	.9368
24	.9150	.9451	.7979
25	.8974	.9515	.9519
26	.9308	.9399	.9276
27	.9278	.9573	.9271
28	.8974	.9482	.9177
29	.8229	.9545	.9154
30	.9249	.9681	.9462

Note. For each of the tapes (sample, main and explanation) the ratings of each untrained rater were correlated with the ratings of each of the remaining 29 raters, resulting in 29 correlations for each rater for each tape. To indicate how well each rater agreed with the other 29 raters, on each of the three tapes the median correlation was determined. Entries in the above Table are these median rater reliabilities.

[a]For the sharing sample tape, 13 behavioral responses were rated. Any $r = .5529$ is significant at $p < .025$; $r = .6835$ is significant at $p < .005$; $r = .8010$ is significant at $p < .0005$.

[b]For the sharing main tape, 42 behavioral responses were rated. Any $r = .3044$ is significant at $p < .025$; $r = .3932$ is significant at $p < .005$; $r = .4896$ is significant at $p < .0005$.

[c]For the sharing explanation tape, 13 behavioral responses were rated. Any $r = .5529$ is significant at $p < .025$; $r = .6835$ is significant at $p < .005$; $r = .8010$ is significant at $p < .0005$.

TABLE IV

MEDIAN CORRELATIONS OF UNTRAINED RATERS

COMFORTING

UNTRAINED RATER	COMFORTING SAMPLE TAPE	COMFORTING MAIN TAPE	COMFORTING EXPLANATION TAPE
1	.8267	.7358	.8226
2	.8186	.7918	.8855
3	.8500	.7797	.7863
4	.8527	.7797	.8457
5	.7545	.7610	.7691
6	.8476	.7836	.8218
7	.7447	.7026	.6101
8	.7329	.6059	.8033
9	.7913	.5372	.7435
10	.8412	.6881	.6750
11	.8861	.6772	.7719
12	.8500	.7803	.8478
13	.8405	.7714	.8750
14	.8822	.6136	.8183
15	.5259	.5685	.8256
16	.8570	.7064	.8707
17	.8871	.7407	.8532
18	.8215	.6856	.8397
19	.6530	.6042	.7689
20	.8611	.7500	.7811
21	.5725	.7421	.8747
22	.8926	.7803	.7103
23	.8926	.7803	.8613
24	.7286	.6742	.8163
25	.8567	.7056	.8324
26	.8324	.7163	.8441
27	.8181	.7472	.8309
28	.8752	.7480	.8277
29	.8352	.7131	.7998
30	.8119	.2971	.7313

Note. For each of the tapes (sample, main and explanation) the ratings of each untrained rater were correlated with the ratings of each of the remaining 29 raters, resulting in 29 correlations for each rater for each tape. To indicate how well each rater agreed with the other 29 raters, on each of the three tapes the median correlation was determined. Entries in the above Table are these median rater reliabilities.

[a]For the comforting sample tape, 12 behavioral responses were rated. Any $r = .5760$ is significant at $p < .025$; $r = .7079$ is significant at $p < .005$; $r = .8233$ is significant at $p < .0005$.

[b]For the comforting main tape, 20 behavioral responses were rated. Any $r = .4438$ is significant at $p < .025$; $r = .5614$ is significant at $p < .005$; $r = .8233$ is significnat at $p < .0005$.

[c]For the comforting explanation tapes, 12 behavioral responses were rated. Any $r = .5760$ is significant at $p < .025$; $r = .7079$ is significant at $p < .005$; $r = .7800$ is significant at $p < .0005$.

For Tables I through IV, any correlation $>.80$ is statistically significant at $p < .0005$; further, from a measurement perspective, a rater reliability of .80 can be considered an acceptable level for research purposes. An examination of Tables I through III revealed that agreement was highest for sharing, with 25 of the 30 raters having all three of their median reliabilities above r = .80. For teaching, the rater reliabilities for 15 of the 25 raters across all three tapes were above r = .80. For helping, 12 of the 25 raters had rater reliabilities above r = .80 across all tapes. Table IV illustrates that the untrained raters tended to agree less on children's comforting responses, with only 17 of the 30 raters agreeing at r = .70 across the sample, main, and explanation tapes. Agreement was somewhat better on the sample and explanation tapes for comforting, where 15 of the 30 raters had rater reliabilities above r = .80.

Interrater Reliability of Trained Raters

The rater reliability obtained by correlating the ratings of the two trained raters for teaching, helping, sharing, and comforting for the sample, main, and explanation tapes are presented in Table V. The correlations of the trained raters' ratings for the main tapes indicated significant agreement ($p < .0005$) between trained raters for all behaviors (i.e., teaching, helping, sharing, and comforting). Trained raters had the highest agreement for sharing (r = .98) and the lowest agreement for comforting (r = .94). Agreements for the trained raters for the sample tapes and thus the explanation tapes as well were also high, with r = .96 for helping, teaching and sharing, and r = 1.00 for comforting.

Comparison: Trained and Untrained Raters

Correlations between the ratings of the average trained rater and the average untrained rater for teaching, helping, sharing and comforting across

TABLE V

RATER RELIABILITIES OF TWO TRAINED RATERS FOR TEACHING, HELPING, SHARING, AND COMFORTING

	PROSOCIAL BEHAVIORS			
	TEACHING	HELPING	SHARING	COMFORTING
SAMPLE TAPE	.98 df=12[a]	.96 df=13	.98 df=11	1.00 df=10
MAIN TAPE	.96 df=40	.95 df=40	.98 df=40	.94 df=18
EXPLANATION TAPE	.98 df=12	.96 df=13	.98 df=11	1.00 df=10

Note. All correlations in this table are significant at $p < .0005$.

[a] Degrees of freedom (N - 2 =df) appear under each appropriate correlation.

the sample, main, and explanation tapes are presented in Table VI (all correlations in this table are significant at $p < .0005$). On the main tapes, agreement between average trained and average untrained raters was highest for sharing ($r = .94$), followed by helping ($r = .93$), teaching ($r = .90$), and comforting ($r = .87$). An examination of Table VI demonstrates that for teaching and comforting, agreement between trained and untrained raters improved over each succeeding tape (i.e., correlations were higher on the main than on the sample tape, and higher on the explanation than on the main tape). Correlations between trained and untrained raters remained the lowest for comforting, ranging from $r = .82$ to $r = .93$.

Table VII presents the means, standard deviations and t-tests of the mean ratings of trained and untrained raters across all children on the main tapes for helping, teaching, sharing, and comforting. For example, to compute the mean data, the ratings for each child on each main tape were averaged across all untrained raters. The same procedure was followed for the ratings of the trained raters for each behavior. These scores were averaged across all children to determine the overall average of the untrained and trained raters for each main tape. Comparison of these averages indicated no significant differences between the trained and untrained raters for teaching, sharing, or comforting. For comforting however, Table VII revealed the means of the trained and untrained raters to be significantly lower than the means for helping, teaching, or sharing, demonstrating a restriction in the range of responses for comforting. A significant difference was found between the average trained and average untrained raters on the main helping tape ($t(41) = 4.55$, p .001). The untrained raters tended to rate children's helping responses higher than did the trained raters. The ratings of the untrained raters were higher particularly when rating between 2.0 and 4.0 on the 5-point ratings scale, as illustrated in Figures 1 through 4.

TABLE VI

PEARSON CORRELATION BETWEEN AVERAGE TRAINED AND AVERAGE UNTRAINED RATERS' RATINGS FOR TEACHING, HELPING, SHARING, AND COMFORTING

	TEACHING	HELPING	SHARING	COMFORTING
SAMPLE TAPE	.89 df=12[a]	.96 df=13	.89 df=11	.82 df=10
MAIN TAPE	.90 df=40	.93 df=40	.94 df=40	.87 df=18
EXPLANATION TAPE	.99 df=12	.94 df=13	.87 df=11	.93 df=10

Note. All correlations in this table are significant at $p < .0005$.

[a]Degrees of freedom (N - 2 = df) appear under each appropriate correlation.

TABLE VII

MEANS, MEDIANS, STANDARD DEVIATIONS, AND T-TESTS FOR TRAINED AND UNTRAINED RATERS ON THE MAIN TAPE FOR ALL BEHAVIORS

	HELPING			TEACHING	
	Trained	Untrained		Trained	Untrained
Mean	2.68	3.05	Mean	2.89	2.90
Median	2.68	2.90	Median	3.50	3.20
Standard Deviation	1.35	1.41	Standard Deviation	1.44	1.23
	t (41) = 4.55 ($p < .001$)			t (41) = .1025	

	SHARING			COMFORTING	
	Trained	Untrained		Trained	Untrained
Mean	2.31	2.43	Mean	1.86	1.92
Median	2.00	1.40	Median	1.00	1.30
Standard Deviation	2.34	1.70	Standard Deviation	1.13	.74
	t (41) = .6365			t (41) = 1.07	

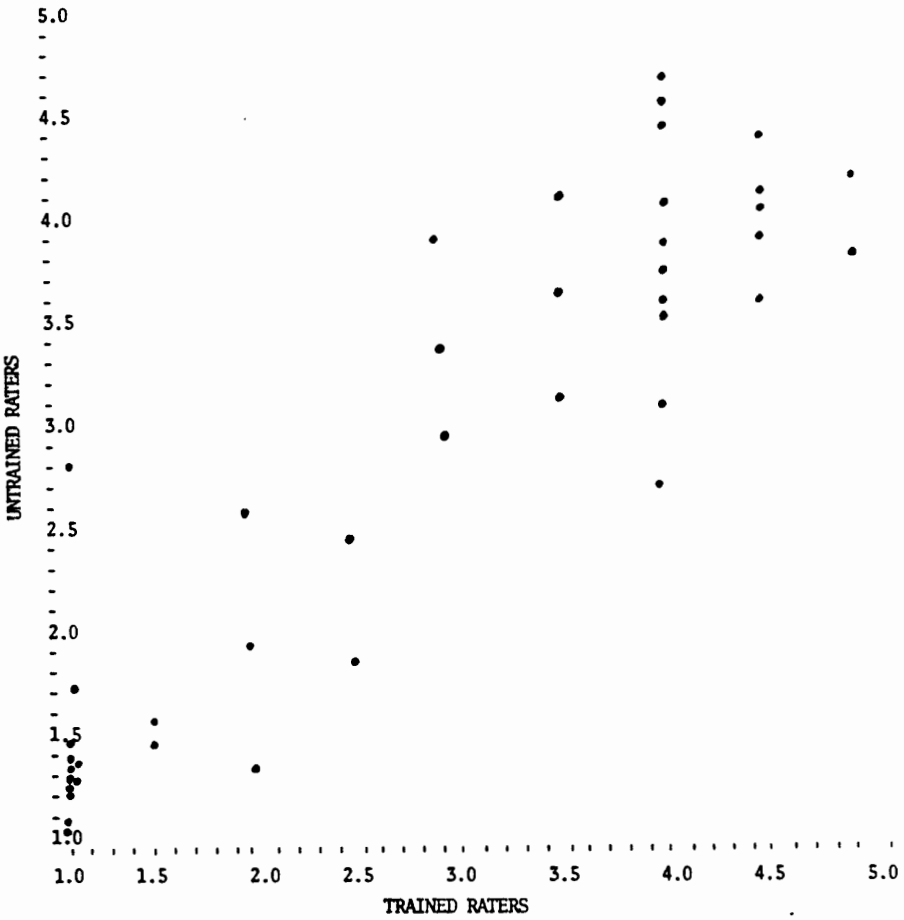

Figure 1. Distribution of average ratings of trained and untrained raters for the teaching main tape.

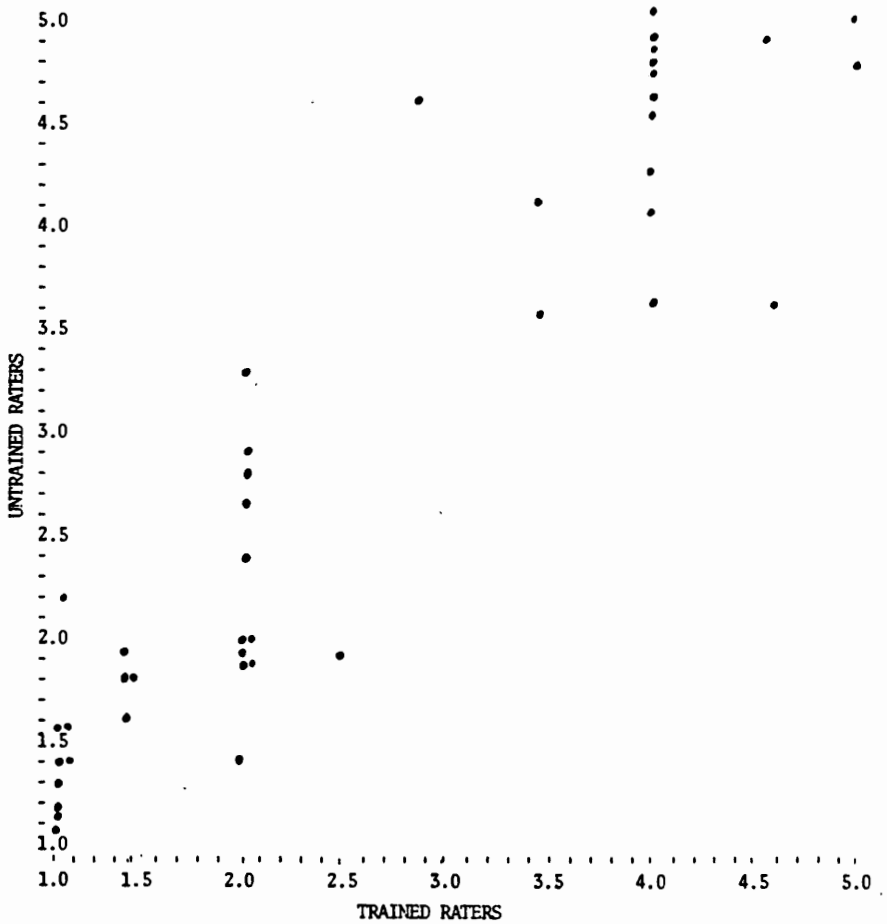

Figure 2. Distribution of average ratings of trained and untrained raters for the helping main tape.

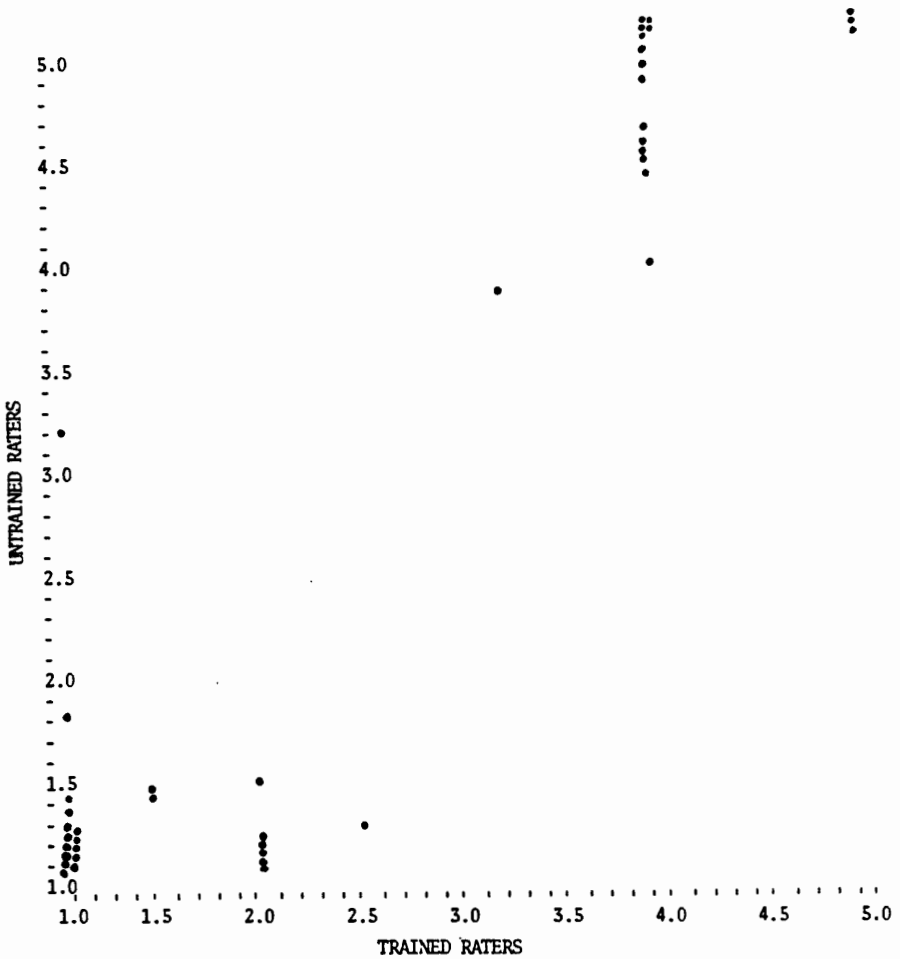

Figure 3. Distribution of average ratings of trained and untrained raters for the sharing main tape.

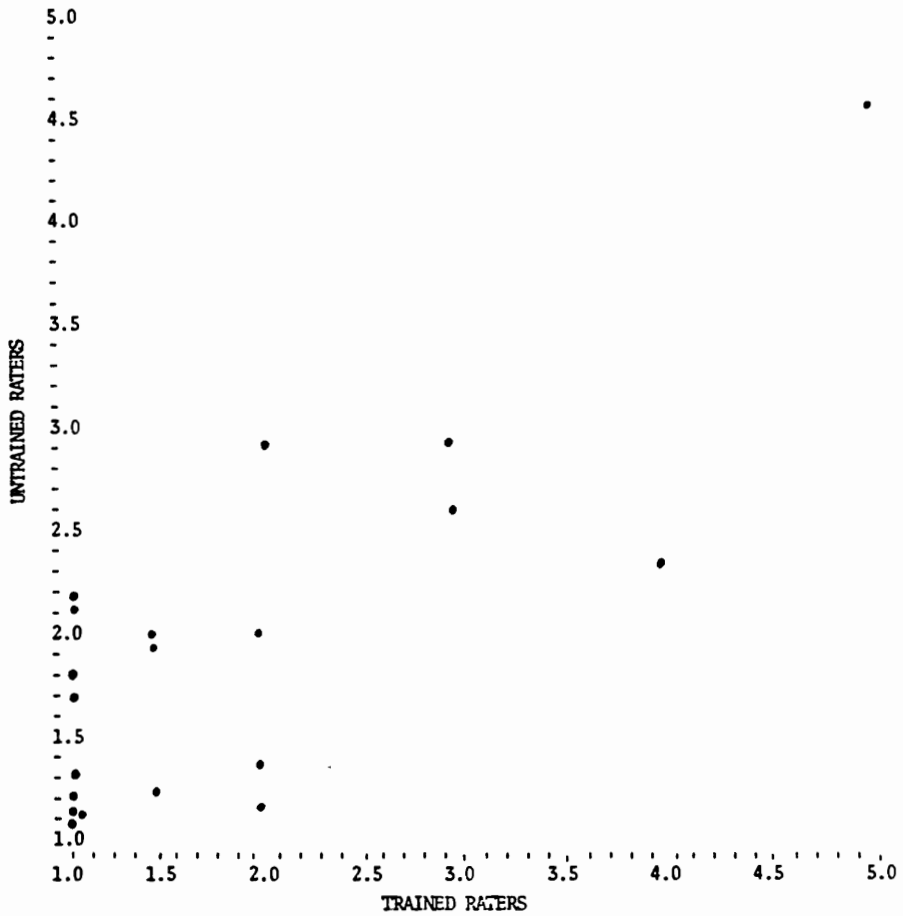

Figure 4. Distribution of average ratings of trained and untrained raters for the comforting main tape.

Content Analysis of Untrained Raters' Explanations, and Comparison with Observation Code Used by Trained Raters

Tables VIII through XI compare the explanations for ratings given by the untrained raters with the observation code used by the trained raters for teaching, helping, sharing, and comforting, respectively. An examination of these tables revealed that the major differences between the explanations given by the untrained raters and the observation code used by the trained raters occurred when untrained raters assigned attributions to the child (i.e., inferences about the child's motivation, intent, or dispositional characteristics). For example, scale point 1 in the observation code used by the trained raters is a "no response" category. However, over all behaviors, an average of 13% of the untrained raters, rather than simply stating at scale point 1 that the child failed to respond went one step further and described the child's lack of response as uncooperative or self-centered. In contrast, other raters interpreted the children's responses by saying that the child wanted to respond but was too young or too shy (6% of the untrained raters for teach, and 4% of the untrained raters for help gave this explanation).

The categories for responses at scale point 2 of the observation code and the explanations given by the untrained raters when assigning a rating at scale point 2 were comparable.

Some overlap in the explanations of untrained raters occurred at scale point 3.0 and 4.0. For teaching, helping, and sharing, untrained raters using scale points 3 and 4 indicated that the child hesitated before responding to the prosocial opportunity. However, this explanation for the child's response occurred three times as frequently when untrained raters used scale point 4 than when they rated at scale point 3. Other explanations of untrained raters for the use of scale point 3 pointed to

TABLE VIII

COMPARISON OF UNTRAINED RATERS' EXPLANATIONS WITH OBSERVATION CODE USED BY TRAINED RATERS: TEACHING

Untrained Raters' Explanations	Scale Point	Observation Code Used By Trained Raters
Child gives no response (64%)[a] Child is uncooperative and ignores adult (14%) Child refuses to teach (9%) Child is self-centered (6%) Child wants to teach but is too young or shy (6%)	1	Child displays no physical or verbal response to the verbalization of need
Child's teaching is incomplete (53%) Child makes suggestion but does not teach (16%) Child shows interest but does not teach (16%) Child physically teaches but gives no explanation (16%)	2	Child teaches only one component of the teaching skill through demonstration or verbal explanation
Child's teaching is partial or incomplete (47%) Child teaches either verbally or physically (27%) Child hesitates before teaching (14%) Child teaches but is half-hearted or feels obligated (12%)	3	Child teaches only two components of the teaching skill through physical demonstration or verbal explanation
Child verbalizes and physically teaches (46%) Child teaches without enthusiasm (35%) Child has good intentions but teaching steps are out of order (20%)	4	Child teaches all three components of the teaching skill through physical demonstration or verbal explanation
Child teaches verbally and physically with eagerness (37%) Child teaches immediately (33%) Child's instructions are very specific (25%) Child teaches and does task for adult (6%)	5	Child demonstrates physically and verbally all components of the teaching skill Child teaches all three components and makes the article for the adult

Note. N = (25 untrained raters rating 14 children's responses)

[a]Numbers in parenthesis indicate the percent of untrained raters using that explanation within each scale point.

TABLE IX

COMPARISON OF UNTRAINED RATERS' EXPLANATIONS WITH OBSERVATION CODE USED BY TRAINED RATERS: HELPING

Untrained Raters' Explanation	Scale Point	Observation Code Used By Trained Raters
Child gives no response (84%)[a] Child is uncooperative or ignores adult (8%) Child wants to but is too young or shy (4%) Child shows interest with no behavior (4%)	1	Child displays no physical or verbal response to the verbalization of need
Child shows inadequate behavior (distracted) (38%) Child shows interest behavior (37%) Child makes suggestions but does not help (25%)	2	Child looks for the lost objects for less than 5 seconds or Child tells the adult someone else will help or Child verbalizes a solution which is unfeasible
Child's helping is incomplete (49%) Child hesitates before helping (24%) Child's helping is half-hearted or the child feels obligated Child makes suggestions but does not help (10%) Child helps but does not give explanation (4%)	3	Child looks for lost objects for more than 5 seconds but does not get out of seat or Child picks up a few of the spilled objects but does not complete task or Child gives a plausible explanation of why the mishap occurred
Child helps the adult but hesitates (45%) Child helps the adult (26%) Child begins helping but does not compelete (18%) Child helps without enthusiasm (11%)	4	Child assists the adult in a task which is better done by more than one person
Child helps immediately (73%) Child helps enthusiastically (29%)	5	Child helps immediately and does everything him/herself

Note. N = (25 untrained raters rating 15 children's responses)

[a]Numbers in parenthesis indicate the percent of untrained raters using that explanation within each scale point.

TABLE X

COMPARISON OF UNTRAINED RATERS' EXPLANATIONS WITH OBSERVATION CODE USED BY TRAINED RATERS: SHARING

Untrained Raters' Explanation	Scale Point	Observation Code Used By Trained Raters
Child gives no response (80%)[a] Child is self-centered (9%) Child is uncooperative (5%) Child refuses to share (5%)	1	Child displays no physical or verbal response to the verbalization of need
Child makes suggestions but does not share (46%) Child shows interest but does not share (43%) Child makes no attempt to share (5%) Child allows adult to use articles but does not share them (5%)	2	Child verbalizes a solution which is unfeasible
Child allows adult to use articles but does not share them (29%) Child felt obligated to share (24%) Child's sharing is partial (19%) Child is concerned but does not share (17%) Child makes suggestions but does not share (12%)	3	Child shares leftovers or a small portion of own materials or Child shares but verbalizes reluctance (i.e., "Now I won't have very much".)
Child shares but hesitates (42%) Child shares part or half (33%) Child shares without enthusiasm (25%)	4	Child relinquishes an object which has been labeled as belonging to the child
Child shares immediately (75%) Child shares with enthusiasm (25%)	5	Child shares all remaining articles

Note. N = (30 untrained raters rating 13 children's responses)

[a]Numbers in parenthesis indicate the percent of untrained raters using that explanation within each scale point.

TABLE IX

COMPARISON OF UNTRAINED RATERS' EXPLANATIONS WITH OBSERVATION CODE USED BY TRAINED RATERS: COMFORTING

Untrained Raters' Explanation	Scale Point	Observation Code Used By Trained Raters
Child gives no comfort (91%)[a] Child is self- oriented (9%)	1	Child gives no sympathetic response or makes a comment which borders on criticism (i.e., "It doesn't hurt me".)
Child looks concerned but does not comfort (80%) Child gives suggestions (20%)	2	Child remembers a similar incident that happened to self others
Child looks concerned (31%) Child gives some verbal concern (28%) Child talks of similar incident (21%) Child gives advice (14%) Child gives physical sympathy (6%)	3	Child asks questions showing concern for the adult
Child gives physical comfort (34%) Child verbalizes comfort (30%) Child talks of similar incident (19%) Child give specific first aid suggestions (13%) Child gives extensive verbal comfort (5%)	4	Child offers comfort or condolence, or expresses concern about another's condition. Child demonstrates physical sympathy
Child gives an intense show of comfort (80%) Child does 3 or more of the following; physical comfort, verbal comfort, advice, talking of similar incident, extensive eye contact (20%)	5	Child verbalizes a sympathetic response and physically demonstrates concern (e.g., kisses the hurt) or Child verbalizes an extensive concern for the adult

Note. N = (30 untrained raters rating 12 children's responses)

[a] Numbers in parenthesis indicate the percent of untrained raters using that explanation within each scale point.

the obligatory or half-hearted nature of the child's response. The exception at scale point 4 between the trained rater's observation code and the untrained rater's explanations occurred when the untrained raters indicated that the child had responded to the prosocial opportunity but had done so without enthusiasm.

At scale point 5, explanations by the untrained raters were comparable to the observation code used by the trained raters. In general, where differences were noted between the untrained raters' explanations and the trained raters' observation code, it was found that the untrained raters had again, assigned attributions regarding the child's motivational or dispositional characteristics (e.g., half-hearted, obligated or self-centered, too shy). When these attributions were made, however, it was not clear from the untrained raters' explanations what it was about the child's behavior that had generated these attributions.

CHAPTER IV

DISCUSSION

The results of this study indicated that untrained raters show a high amount of agreement among themselves when given minimal guidelines for rating the prosocial responses of young children. Further, the ratings of these untrained raters agree substantially with the ratings of raters trained in the systematic observation of children's prosocial responses. And finally, with a few exceptions, untrained raters appear to assign the same reasoning for their rating choices as was apparent in Smith's (unpublished research) observation code derived from the responses of a large number of children in standardized prosocial opportunities.

In general, untrained raters showed substantial agreement with each other when rating the prosocial responses of teaching, helping, and sharing. Agreement was highest for the behavior of sharing, perhaps because of the relative clarity in children's responses to sharing opportunities, i.e., children have a particular item to share and except for partial sharing, they tend to share all or nothing. This sharing then is easily observable.

Untrained raters showed somewhat less agreement among themselves in their ratings of teaching and helping. This is perhaps due to the variety of responses available to a child who is faced with an opportunity to teach or help. For example, children can teach the adult all components of the teaching skill, or only one or two of the components. They can verbalize or demonstrate these skills for the adult, or do both. In cases where the child demonstrates only, it is sometimes difficult to determine

whether the child is merely continuing the teaching activity or is actually "showing" the adult how to do it. All of these factors combine to make the teaching responses of children less clear. Helping opportunities are similar in that they also may have several outcomes. For instance, when picking up the spilled sticks a child may pick up all the sticks, part of the sticks, show concern for the spilled sticks, or offer a suggestion for helping the adult.

Raters showed the least amount of agreement with each other in rating comforting responses. Comforting responses are many and varied in the general population, especially among young children, who perhaps have had more experience being comforted than in comforting others. Thus, it may be more difficult for untrained raters to rate children's comforting responses when they have not had extensive experience with the ways young children respond to situations in which they are asked to comfort an adult. In addition, since the main comforting tapes contained approximately one-half (20 instead of 42) of children's responses as was contained in the main tapes for teaching, helping and sharing, there were less opportunities for the untrained raters to rate. A restriction in the range of children's responses (i.e., children's comforting responses contain a high incidence of scale point 1 and 2 responses) may have also contributed to the somewhat lower reliability among the untrained raters.

When the ratings of the untrained raters were compared to those of the trained raters, agreement was high across all behaviors (teaching, sharing, helping, and comforting) for all tapes (sample, main, and explanation). Agreement between the trained and untrained raters was highest for the behavior of sharing, followed by helping, teaching, and comforting. This substantial agreement between the ratings of the trained and untrained raters demonstrated that untrained raters, after a short opportunity to view the

ranges of responses to prosocial opportunities (i.e., the sample tapes), can rate children's responses as effectively and in the same manner as do trained raters.

As with the comparisons among the untrained raters, agreement was highest between the trained and the untrained raters for the behavior of sharing. Again, this is perhaps due to the concrete quality of children's responses to sharing opportunities (i.e., the child shares a cup or doesn't, the child shares all or part of a cookie, eats the cookie him/herself, or makes a suggestion for how the adult can get what she wants). Agreement between the trained and untrained raters, while still substantial, was lowest for comforting. It may be more difficult for untrained raters, who have had relatively little experience in the ways in which children comfort, to agree with the ratings of trained raters experienced in assessing children's comforting responses. Again, the reduced number of prosocial responses available to the raters, and the restriction of range (i.e., the high incidence of responses at scale points 1 and 2) may have also contributed to the somewhat lower interrater reliability.

There was some improvement in raters' agreement over the sample, main, and explanation tapes for teaching and comforting. It is possible that untrained raters, after viewing a larger sample of teaching responses, would improve their ability to determine when a child was merely continuing the activity of planting seeds or was actually demonstrating the task for the adult. With practice, raters may have become more attuned to the subtle indications of children's teaching attempts. For example, they may have begun to realize that a child who was silently demonstrating the teaching skill, looked at the adult occasionally to see that the adult had received the instructions, or that a child made a comment when the teaching was concluded, such as "That's it". In rating comforting responses, untrained raters

may have begun rating children's responses by comparing them to the ways in which adults comfort another person. After some experience in viewing children's comforting responses, it is possible that untrained raters began to more familiar with the ways in which children comfort an adult. Under these conditions an improvement between the ratings of the trained and untrained raters over the sample, main, and explanation tapes would be expected.

Quantitative differences between the ratings for the trained and untrained raters were seen when the untrained raters used higher scale points than trained raters in rating children's helping responses. An analysis of explanations given by the untrained raters at scale points 3 and 4 indicated that raters frequently assigned attributions to the children's responses, thereby increasing their ratings. For instance, a partial helping response was rated at scale point 3 by the trained raters, while the untrained raters stated that the child had "good intentions" and rated the partial helping at scale point 4. An example of a sequence in which the untrained raters cited the child's "good intentions" to help occurred when a child got up immediately to pick up the spilled sticks, but once on the floor became preoccupied with looking at the flower stickers and did not complete the task. The effect of these attributions, while slightly increasing the differences between the ratings of the trained and untrained raters, was not significant except for the behavior of helping.

A qualitative examination of the explanations generated by the untrained raters revealed that, in general, untrained raters used the same rationale in assigning scale points to children's prosocial responses as did the trained raters using the observation code. The explanations for ratings given by the untrained raters for each scale point contained the same rationale as did the categories in the observation code used by the trained

raters. However, the untrained raters went one step further and assigned various dispositional and motivational characteristics to the responses of some of the children. For example, at scale point 1, in addition to a "no response" category, untrained raters described some children's behavior as uncooperative or self-centered. Additionally, some children were described as wanting to respond but being unable to because they were too young or too shy. This type of attributional explanation for children's prosocial responses was also used at scale points 3 and 4, where untrained raters stated that some children responded half-heartedly or felt obligated to behave prosocially. At scale points 2 and 5, no additional attributions were assigned to the children's responses and the explanations given by the untrained raters were comparable to the rationale contained in the observation code used by the trained raters.

The assignment of attributions to some children's responses, then, appears to be the major difference between the explanations provided by the untrained raters and the observation code used by the trained raters. This tendency by the untrained raters to assign motivational or dispositional attributions to children's prosocial responses is consistent with what we know about the way people make sense of the behavior of others in everyday life. From a scientific standpoint, however, it is apparent that these explanations for the behavior of others are not always accurate. For this reason, when an observation code is constructed, only the actual physical or verbal responses are recorded and no speculations as to the child's motivation or personality characteristics are made, as suggested by Sackett (1978). However, while the untrained raters assigned attributions to the children's prosocial responses which were not included in the observation code used by the trained raters, this tendency did not significantly affect the correlations between the ratings of the trained

and untrained raters. In fact it should be noted that the high degree of similarity between the untrained rater's explanations and the observation code used by the trained raters, coupled with the significant correlations between the ratings of the trained and untrained raters, clearly demonstrates that the observation code and rating scale developed by Smith (unpublished research) meets the empirical criterion for construct validity. That is, the results obtained from one measure of prosocial behavior (i.e., the explanations and ratings of the untrained raters) were much the same as those obtained for the other measure (i.e., the observation code and ratings of the trained raters). Therefore, this observation code and rating scale can be thought of as having a high degree of construct validity. Given these findings it is possible to speculate that an observation code may allow for additional attributional inferences without sacrificing reliability or validity.

The explanations provided by the untrained raters indicated that untrained raters, and perhaps people in general, view children's responses to opportunities to help, teach, share, and comfort in nearly the same way as do raters trained in the systematic observation of children's prosocial responses. If this is so, then it can be inferred from these findings that the observation code and rating scale described earlier are consistent with the way people in general view children's responses to opportunities to help, teach, share, and comfort. The agreement between the explanations provided by the untrained raters and the guidelines contained in the observation code indicate that this observation code and rating scale meet the criterion for content validity, i.e., that a measurement instrument measures what it was designed to measure.

In summary, the results of this study suggest that the observation

code and ratings scale developed by Smith (unpublished research) is a reliable and valid instrument for the measurement of prosocial behavior in young children. When given minimal guidelines untrained raters showed a significant amount of agreement among themselves when rating prosocial responses of young children. In addition, and most importantly, the ratings of untrained raters were correlated to a significant degree with judgments of persons skilled in the use of systematic observational methods. And finally, the reasons given by the untrained raters for their ratings were comparable to the rationale contained in the observation code and ratings scale used by the trained raters. It would seem, therefore, that this observation code and rating scale could be used by other researchers confidently to measure prosocial behavior in young children.

REFERENCE NOTES

Smith, C. L. Training prosocial behavior in preschool children. Paper presented at the Western Psychological Association Convention, San Francisco, 1978.

Blackwell, J. M., Smith, C. L., & Stewart, B. J. The effects of instructions on prosocial behavior of preschool children. Paper presented at the Western Psychological Association Convention, Hawaii, 1980.

REFERENCE LIST

Cronbach, L. J. Essentials of psychological testing. New York: Harper and Row, 1960.

Nunnally, J. C. Psychometric theory. New York: McGraw Hill, 1967.

Sackett, G. P. Observing behavior. Volume II. Data collection and analysis methods. Baltimore: University Park Press, 1978.

Staub, E. To rear a prosocial child. In D. J. DePalma and J. M. Folly, (Eds.) Moral development: Current theory and research. Hillsdale, N. J.: Lawrence Erlbaum, 1975.

Yarrow, M. R., Scott, P. M., and Waxler, C. Z. Learning concern for others. Developmental Psychology, 1973, 8, 240-260.

Yarrow, M. R., and Waxler, C. Z. Dimensions and correlates of prosocial behavior. Child Development, 1976, 47, 118-125.

APPENDIX

OBSERVATION CODE AND RATING SCALE FOR PROSOCIAL BEHAVIOR

SCORE	CODE	BEHAVIOR	
4	H	Helping:	Child assists in a task which is better or more quickly done by more than one person, e.g., finding a lost object (child must get up from seat and look for object for at least 5 seconds); locating any needed object; picking up objects which have dropped to the floor; moving objects from one place to another; clearing objects or materials from table. If child looks for lost object for 5 seconds or more but does not get out of seat, score 3. If child looks for lost object for less than 5 seconds, score 2. If child helps within one second and does everything by him/herself (e.g., picking up all sticks alone), score 5.
4	S	Sharing:	Child relinquishes an object which had been in the child's possession or use, or which was owned by the child (owernship must be previously established by telling the child, "This is yours"). If child shares only leftovers (e.g., playdough scraps not in shape of cookie), or very small portions of own materials, score 3. If child shares all remaining materials, score 5.

SCORE	CODE	BEHAVIOR	
4	T	Teaching:	Child instructs another in a skill or activity. The instruction can be through physical demonstration or verbal explanation as long as the child gives another information which enables the individual to continue or complete an activity. All 3 components of teaching responses must be demonstrated or explained. If child teaches only 2 components, score 3. If child teaches only 1 component, score 2. If child demonstrates physically and verbalizes all 3 components of teaching response, score 5.
4	Sym	Sympathy:	Child offers comfort or condolence, or expresses concern about another's condition. Verbalizations must include words such as sorry, hurt, better, alright, okay, etc. Verbalizations scored as sympathy include: -"It's alright." -"Sorry, I know you're hurt" or "I bet it hurts." -"It's okay" or "That's okay" or "It will be okay." -"I think it will stop hurting now." -"It will feel better in awhile" or "It'll get better." -"I'm sorry." -"I wish it didn't hurt." -"Have to get a bandaid for you so it won't hurt!" -"Are you alright?" -"Does it feel better?" Physical demonstrations of comfort or sympathy include extending a hand or arm toward the injured person and patting, stroking, hugging, kissing in a positive manner. Physical demonstrations receive a score of 4. If a child displays negative effect, and/or repeats the verbalization of need or equivalent (e.g., "Ouch!") score 2. Note: affect is

(Appendix Continued)

SCORE	CODE	BEHAVIOR	
			scored only when there is no physical or verbal response. If child remembers a similar past incident or event which happended to self or others (e.g., "I got an owie and it bleeded"; "I hurted myself once"), score 2. If child's statement lacks sympathy or condolence or borders on criticism, or includes an account of own coping behavior in similar situations (e.g., "When I touched it, it didn't hurt me!"; "You didn't hit it very hard"; "What did you do that for?"; "That's what you get"), score as 1. If child verbalizes a sympathetic response (e.g., "I'm sorry") and displays another prosocial response (e.g., helping or sharing) at level 4, score as 5. If child verbalizes a sympathetic response and physically demonstrates a response, e.g., kisses the hurt, score as 5. If the child verbalizes an extensive sympathetic response (e.g., "I'm sorry you hurt yourself. It will get better soon"; "Want to put something on it? I believe it does really hurt. It will heal. I don't think it will be a bruise"), score as 5.
3	R_3	Remedy 3:	Any neutral or positive verbal response by the child which poses a solution to the problem implied by the verbalization of need. The following are examples of Remedy 3: -(lost pen) "But we could go outside where you were. Could write with the brush." -(no cup) "I'm going to bring one for you cause I didn't bring one." -(no cup) "You could get a different one. Use that glass one."

(Appendix Continued)

SCORE	CODE	BEHAVIOR	
			-(no cup) "You can plant in a garden. I planted in a garden with my dad." -(no cup) "You can have that one (pointing to model)." -(no flower sticker) "Take one off there (off model)." -(no flower sticker) "Do you want a stem? Here's stem. Somebody must have pulled off the flower." -(no snack) "Do you have some at home? Buy some. Are you going to buy some?" -(no glitter) "Do you have some at home? Why don't you use some at home?" -(bumped head) "Go out there and get a cold pack then." -(bumped head) "Maybe we have some bandaids" or "Do you need a bandaid?"
2	R_3	Remedy 2:	Any neutral or positive verbal response by the child which falls into one of the following categories: a) Child tells adult to engage in the behavior herself, e.g., -(no cup) "Get one can't you? Aren't you allowed to get one yourself?" -(things need to be moved to the table) "Alright--do that." -(spilled sticks) "Pick 'em up." -(spilled sticks) "Well, you'll have to pick them up." -(spilled sticks) "You pick them up because you spilled them." -(lost box) "Go look for it" or "Look on the floor." b) Child says that someone else (e.g., the other adult) engage in the behavior, e.g., -(no flower sticker) "She's gonna go get some." -(lost box) "Ask the other girl when she comes back." -(no snack) "She'll give you one."

(Appendix Continued)

SCORE	CODE	BEHAVIOR
		-(no snack) "Well, she can go get some more." -(spilled cookie cutters) "She'll do it."
	c)	Child "admonishes" adult by offering comments regularly made by socializing agents in similar situations, e.g., -(spilled sticks) "You shouldn't have dropped them like that." -(spilled sticks) "That's cause you shouldn't have opened it." -(spilled sticks) "Don't drop them again." -(stubbed toe) "You better watch where you're going." -(stubbed toe) "What's there? You didn't see that." -(don't know how to plant seeds) "You could if you wanted to." -(don't know how to water seeds) "I wanna see if you can." -(don't know how to water seeds) "Well, you have to try." -(bumped head) "You better be careful." -(lost box) "Where'd you put it? Stand there till you remember." -(spilled papers) "You better be careful."
	d)	Child makes an observation concerning the constraints within the situation, e.g., -(no cup) "I know--you missed all of it." -(no glitter) "Where is it? This is for me." -(no stars) "These are the only ones." -(no snack) "There's only three--cause I like animal crackers." -(no snack) "Only three for me." -(don't know how to water seeds) "There's no more cups." -(don't know how to do a flower sticker) "She just took the sticks away."
	e)	Child offers an explanation for the adult's state of need, e.g.,

SCORE	CODE	BEHAVIOR	
			-(no cup) "Where are the cups? Someone stole them?" -(lost box) "Where'd you put it?" -(lost pen) "Maybe it went to your home."
2	VP	Verbal Post-ponement:	Child promises to behave prosocially at a later time, but does not follow through. -"I'll do it for you later." -"Just a minute." -"I'll find it after I'm done." -"I'll show you when I get through."
1	A_1	Associa-tion:	Child talks about content of need verbalization without apparent recognition of the need. -(don't have any cookies) "One time when my mom and I went to the movie we bought this kind of cookies." -(don't know how to cut a cookie) "I'm gonna make a ball." -(don't know how to plant seeds) "I have two cups of dirt now."
1	ACK	Acknow-ledgment	Child verbally demonstrates awareness of another's need, e.g., by repeating or paraphrasing need verbalization. -(no cookies) "There's none for you." -(no cookies) "You got no cookie." -(no playdough) "You don't have any." -"Oh." -"Uh-huh." -"It did?" -"I do." -"Uh, oh." -(no seeds) "There's no seeds in there either."
1	D	Diver-sion:	A verbal response by the child about an unrelated topic.

SCORE	CODE	BEHAVIOR	
1	NR	No Response:	Child displays no physical or verbal response to verbalization of need.

Additional Scoring Rules
(General)

1. Subject's responses which are delayed (occur after 7 seconds following need verbalization, model, or prompt) receive a score of 1 point less.

2. Subjects who respond prosocially but verbalize reluctance and/or reasons why s/he shouldn't (e.g., "Now I won't have very much"), score as 1 point less.

3. Subjects who report the inequity either before (e.g., "What about the other girl?") or after the need verbalization (e.g., "She doesn't have any playdough"; "She couldn't find her special box"), or who display a continued recognition of the need (e.g., "She bumped her head"; "Do you know where the special box is?") receive a score of 1 point more.

4. Subjects who respond prosocially before the need is verbalized receive a score of 1 point more.