Reports from the Research Laboratories

of the

Department of Psychiatry

University of Minnesota

The Effect of the Behavior Base-rate on the Agreement Between Two Raters of the Behavior

by

ROBERT R. GOLDEN and GORDON T. HEISTAD

Report Number PR-77-1

January 1977

Reports from the Research Laboratories

of the

Department of Psychiatry

University of Minnesota

The Effect of the Behavior Base-rate on the Agreement Between Two Raters of the Behavior[1]

by

Robert R. Golden and Gordon T. Heistad

Report Number PR-77-1 January 1977

[1]This research was supported in part by grants from the Psychiatry Research Fund, the University of Minnesota Graduate School and the University Computer Center.

Abstract

A model is developed for measurement of the agreement between two raters of the presence or absence of a behavior. Internal validity tests of the model are developed and tested by Monte Carlo trials. In an application to certain behavior ratings of psychiatric patients on over 24,000 occasions it is found that the disagreement between the raters is mainly a function of the behavior base-rate.

The Effect of the Behavior Base-rate on the Agreement Between Two Raters of the Behavior

The Problem

Suppose two raters observe a number of individuals for the presence or absence of some behavior or complex of behaviors. Denote the presence of the behavior by + and its absence by -. In one particular example their agreement may be given in terms of the proportions of a four-fold table such as in Table 1. In this example, the phi-coefficient is

$$\frac{.77 - (.86)(.86)}{\sqrt{(.86)(.14)(.86)(.14)}} = .25$$

and the kappa-coefficient (see Cohen, 1960) is

$$\frac{.77 - (.86)(.86)}{1 - (.86)(.86)} = .12.$$

If the behavior is perceived by one rater, the probability it will also be perceived by the other rater is $.10/(.18 + .10) = .36$. This latter quantity will be referred to as the "proportion agreement rate for the + ratings." From each of these three values, one is led to conclude that rater agreement or reliability is rather low. We can then further conclude by an often used psychometric bromide, that if the reliability is low then the validity in the perception of the behavior is also low. In the example, then, this validity must be near nil.

The point of this report is to show that the above reasoning

is too simplistic, and is incomplete if not simply incorrect. In the particular example above, it turns out that the rater validity is as high as we might reasonably hope for under general plausible conditions.

The substantive problem that gave rise to this methodological problem was that of detecting behavioral changes in hospitalized patients when a change in behavioral or psychotropic drug treatments takes place. To rate behaviors a particular behavior coding system was used. Jones, Reid, and Patterson (1975) have summarized the results of ten years of research with a 28-item behavior coding system used by investigators at the Oregon Research Institute to quantify the frequency of children's behavior problems based on home observations by trained observers. Their data indicate quite acceptable overall reliability (inter-observer agreement), and sensitivity (ability to discriminate improvement after treatment) for their instrument and technique. Attention was given to meet the special needs of research in a seriously retarded population in an institutional setting. That code was expanded to the 31 items shown in Table 2. Some of the definitions of code words were altered to make them more suitable for the special population and settings. Then two trained observers coded the behavior of ten patients with serious, but highly variable, behavior problems for approximately 25 three-minute time samples per week for seven weeks. During each time sample, both observers coded the behavior of the same patient. Every six seconds they recorded the occurrence of

at least one and a maximum of two behaviors of that patient from the list of 31 behavior codes. They also recorded, for each 6-second observation, the identity of any other person (staff or patient) who might be interacting with the target patient under observation and recorded one behavior of that other person whenever inter-personal interactions occurred during the observation sessions.

The purpose of the pilot study reported here was to evaluate the coding system for use in a research project which will be aimed at determining both the favorable and the unfavorable behavioral effects of drugs for individual patients and for certain classes of patients. The principle objective will be to evaluate the effectiveness of continued tranquilizer medication in the control of aggressive, hyperactive, and other maladaptive behavior in retarded patients who have been previously maintained on such drugs for more than six months. In order to do this, changes in behavior when the medication is (blindly) discontinued will be observed.

Method

For analysis of the data, a psychometric model was developed. Let us use the following notation:

p_b: the probability that a given rater will correctly detect the behavior if, in fact, it is present,

p_n: the same probability when, in fact, the behavior is not present,

P: the proportion of the observations when the behavior actually is present,

Q: the proportion of the observations when the behavior actually is not present (= 1-P).

The manifest four-fold proportions are given in Table 3. We assume now (to be checked later) the following:

A_1: The behavior is either present or absent in each observation,

A_2: The parameter p_b is the same for each rater, and p_n is the same for each rater.

A_3: The raters make uncorrelated or independent errors in the ratings of the same behaviors.

The first assumption (A_1) will be violated, for example, if the behaviors are not defined such as to be of an all or none character but are, in fact, a matter of degree or if more than one behavior occur simultaneously. We allowed for the latter possibility by allowing the raters to code either 1 or 2 behaviors for each 6 second interval but they coded 2 behaviors less than 5 percent of the time. The second assumption (A_2) is violated if one rater is more accurate than the other; it is required here purely for mathematical reasons. The third assumption (A_3) is violated if, for example, the raters are influenced by the same biases, distractions, misinterpretations and the like. We will have more to say about these assumptions later so suffice it to say here that we only wish to determine if such an idealized model is approximately consistent with the data.

It follows directly from the assumptions that (where $q = 1-p$ and $Q = 1-P$)

$$p_b^2 P - p_n^2 Q = p^{++} \tag{1}$$

$$p_b q_b P + p_n q_n Q = p^{+-} = p^{-+} \tag{2}$$

$$q_b^2 P + q_n^2 Q = p^{--} \tag{3}$$

for each of the four cells of Table 3. It should be noted that

(a) A_2 and A_3 imply $p^{+-} = p^{-+}$ and $p_1^+ = p_2^+$,

(b) only two of the equations (1), (2), and (3) are independent, and

(c) there are three unknowns P, p_b and p_n.

We can solve the above three equations if we assume that P is quite small (say, for the moment, less than .15). That is, we will develop this model for the situation where the base-rate of the behavior is sufficiently small. If the approximate size of the base-rate is not known a priori then we will have to rely on some sort of tests to tell us if it is small enough or not for the model to give accurate results. If $P \doteq 0$, we have from (2)

$$p_n q_n \doteq p^{+-} \text{ or } p_n q_n \doteq p^{-+}$$

or

$$p_n^2 - p_n + p_a \doteq 0, \tag{4}$$

where $p_a = \frac{1}{2}(p^{+-} + p^{-+})$;

hence,

$$\hat{p}_n = \frac{1 - \sqrt{1-4p_a}}{2}, \qquad (5)$$

where the carat denotes the estimator of the parameter.

Likewise, from (3) we have

$$\hat{Q} \doteq \frac{\bar{p}^{--}}{q_n^2} = \frac{4\bar{p}^{--}}{(1 + \sqrt{1-4p_a})^2}. \qquad (6)$$

For the data of the example given above, from (5) we obtain

$$\hat{p}_n = \frac{1 - \sqrt{1-4(.09)}}{2} = .10,$$

and from (6)

$$\hat{Q} = \frac{.77}{(.9)^2} = .951, \quad \text{or} \quad \hat{P} = .049.$$

It follows from (1) that $\hat{p}_b = .90$. Thus we see that if the model is valid the raters in the example correctly detect both the presence and absence of the behavior with 90% accuracy ($\hat{p}_b = .90$, $1 - \hat{p}_n = .90$). The low agreement rate with regard to the presence of the behavior is due mainly, almost solely, to the low base-rate of the presence of the behavior.

The corpus of a psychometric model, the parameter estimation procedures, cam be complemented by "consistency tests" (the term was first used by Meehl, Note 1), the results of which describe how well the model fits the data. The major purposes of consistency tests are to avoid spurious findings (e.g., the model indicates the behavior is dichotomous-like when, in fact, it is nothing of the sort) and to detect when

the parameter estimates are too erroneous for the particular purposes of the study. In brief they check the verisimilitude or truthlikeness of the model assumptions. Consistency testing in connection with a psychometric model is just as obviously required, as much a matter of simple common sense, and as easily done as in many other similar endeavors. For example, the builders of the Minnesota Multiphasic Personality Inventory (MMPI) realized that validity keys were required. While anyone would know that some people randomly respond, lie and so on when taking the MMPI, it is curious that few psychologists or sociologists act as if Nature could, on occasion, be more devious than mathematicians require. What is needed are validity scales for psychometric models. Such scales can be derived from the assumptions of the model.

The assumptions of the model probably never have perfect verisimilitude; in other words, there usually is some amount of "assumption departure." But, of course, it is only necessary that the assumption departure be adequately small, or, to put it another way, that assumptions be adequately "robust," so that parameter estimates are accurate enough and so that the possibility of spurious results can be dismissed. It will be shown that for the present model that the extent of assumption departure can be checked by the use of statistical tests which are derived in part from the model assumptions, and in part from Monte Carlo trials of the model.

A psychometric model, of the type under consideration in this paper, can be thought of as a set of equations relating

a set of latent (unknown) parameters (P, p_b, p_n) to a set of manifest parameters (the four-fold proportions). While some of these equations may involve only latent parameters and others involve only manifest parameters, the equations of the most immediate concern in the development of the model involve both kinds of parameters.

There are two special types of these equations used in a psychometric model: (a) those that state the assumptions and (b) those equations derived from the assumptions which express the latent parameters as explicit functions of the manifest parameters (hence referred to as the "estimation equations"). However, there remains further mathematical derivation to better prepare the model for application to substantive problems. Such derivation can be very roughly described to be that of deriving all the different "interesting" relations between the parameters (latent and manifest) that one is able to. Hopefully, these latter equations or conditions can then be used for determining how well the model fits the data; hence they are called the "consistency equations or relations." If the assumptions are roughly correct and the estimates of the manifest parameters (obtained directly from the data, of course) and of the latent parameters (by the calculations given by the estimation equations) are roughly correct, then the parameter estimates will roughly satisfy the consistency equations.

If the consistency equations and the corresponding tests are developed to flow from the assumptions so as to provide

sensitive checks of the verisimilitude of the assumptions, then the situation where the parameter estimates adequately satisfy the consistency equations should usually only obtain if the results are, in fact, accurate and not spurious.

Since the assumptions A_1, A_2, A_3 are not perfectly true, and the above solutions of equations (1), (2) and (3) are exact only when $P = 0$, there will typically be errors in $\hat{p}_b$, $\hat{p}_n$, and $\hat{P}$. Assume, for the moment, that there is no sampling error. Since the difference between the estimate of each parameter and its true (but unknown and latent) value is usually not zero; that is, the quantities δP, δp_b, δp_n, as defined below, are (usually) non-zero estimate errors:

$$\delta P = \hat{P} - P$$

$$\delta p_b = \hat{p}_b - p_b$$

$$\delta p_n = \hat{p}_n - p_n$$

These estimation errors result in equations (1), (2), and (3) not being <u>perfectly</u> satisfied by the set of parameter <u>estimates</u>. Presumably, if the estimates are grossly in error, the departures between the left and right sides of each of (1), (2), and (3) will be relatively large.

The amount of such departure in the three equations which is allowable can be approximated by the use of the exact differential. For example, (1) is of the form

$$p^{++} = F(p_b, p_n, P);$$

hence

$$p^{++} = \frac{\partial F}{\partial p_b} dp_b + \frac{\partial F}{\partial p_n} dp_n + \frac{\partial F}{\partial P} dP,$$

or

$$dp^{++} = 2p_s P dp_s + 2p_n Q dp_n + (p_s^2 - p_n^2) dP.$$

It follows that for sufficiently small estimate errors, δp_b, δp_n, and δP, that

$$\delta p^{++} \doteq 2p_b P \delta p_b + 2p_n Q \delta p_n + (p_b^2 - p_n^2)\ \delta P,$$

where

$$\delta p^{++} = p^{++} - F(\hat{p}_b, \hat{p}_n, \hat{P}) = p^{++} - (\hat{p}_b^2 P + \hat{p}_n^2 \hat{Q}).$$

Thus, we have the manifest departure of p^{++} from $F(\hat{p}_b, \hat{p}_n, \hat{P})$ as a function of the estimates $\hat{p}_b, \hat{p}_n$, and $\hat{P}$ and the latent errors of estimate δp_b, δp_n, δP. Suppose we require that these errors of estimate not exceed, in absolute magnitude, a value of .10. Such a value was first thought to be a reasonable requirement for such a problem as the present one. Since it can be shown by Monte Carlo study that it is nearly always true that $\delta p_b > 0$, $\delta p_n < 0$, and $\delta P > 0$ it follows that it is reasonable to require that

$$\left|\delta p^{++}\right| \leqq .10 \left|2\hat{p}_b \hat{P} - 2\hat{p}_n \hat{Q} + \hat{p}_b^2 - \hat{p}_n^2\right|. \qquad (7)$$

If the above condition is not met then it is evidently true that the postulated behavioral situation is not sufficiently close to the actual one. Thus, Condition (7) will thus be

tried as a consistency test below.

Likewise, from (2) we obtain

$$\left|\delta p^{+-}\right| \leqq .10 \left|\hat{p}_n\hat{q}_n - \hat{p}_b\hat{q}_b + (1-2\hat{p}_b)\hat{P} - (1-2\hat{p}_n)\hat{Q}\right| . \quad (8)$$

The same condition as in (8) obtains for $\left|\delta p^{-+}\right|$. Finally, from (3) it follows that

$$\left|\delta p^{--}\right| \leqq .10 \left|\hat{q}_n^2 - \hat{q}_b^2 - 2\hat{q}_b\hat{P} + 2\hat{q}_n\hat{Q}\right| . \quad (9)$$

Along with the above four consistency tests we can require that

$$\hat{p}_b > \hat{q}_b \quad \text{and} \quad \hat{q}_n > \hat{p}_n. \quad (10)$$

Condition (10) may be useful in detecting grossly invalid ratings and non-dichotomous behavior.

In similar vein, we can require that

$$\left|p_{b_1} - p_{b_2}\right| < .10 \quad \text{and} \quad \left|p_{n_1} - p_{n_2}\right| < .10.$$

The assumption A_2 formally requires these differences to be zero but the above inequalities are acceptable requirements concerning the robustness of the method with respect to A_2. Since $p_1^+ - p_2^+ = (p_{b_1} - p_{b_2})P + (p_{n_1} - p_{n_2})Q$, it follows that

$$\left|p_1^+ - p_2^+\right| \leqq .10. \quad (11)$$

Preliminary Monte Carlo study has shown that the above tests do not adequately detect excessive correlation between ratings either when the behavior is present or when it is absent. If we know that it is likely that $p_b \leqq .9$, $p_n \geqq .10$, $q_b \geqq .10$, and $q_n \leqq .9$, then under A_3 it is easily shown that we must have

$$\phi \leqq .50. \tag{12}$$

Also, Monte Carlo study has shown that for the method to work adequately it should be required that

$$\hat{P} \leqq .15. \tag{13}$$

The above eight consistency test requirements (7) — (13) were used in a Monte Carlo study of the method. The base-rate P was assigned the values .05, .10, and .15, p_{b_1} and p_{b_2} were assigned the values .5, .6, .7, .8, and .9, p_{n_1} and p_{n_2} were assigned the values .1, .2, .3, .4, and .5 and finally the taxonic class covariances c_s and c_n were each assigned values of 0, .05, and .10. Using every combination of these parameter values (there are 3x5x5x5x5x3x3 = 16,875 different combinations), the proportions of the four-fold table were determined by the following identities:

$$p^{++} = (p_{b_1} p_{b_2} + c_b)P + (p_{n_1} p_{n_2} + c_n)Q$$

$$p^{+-} = (p_{b_1} q_{b_2} + c_b)P + (p_{n_1} q_{n_2} + c_n)Q$$

$$p^{-+} = (q_{b_1} p_{b_1} + c_b)P + (q_{n_2} p_{n_2} + c_n)Q$$

$$p^{--} = (q_{b_1} q_{b_2} + c_b)P + (q_{n_1} q_{n_2} + c_n)Q$$

Then the parameter estimation equations and the consistency tests were applied to each of the 16,875 sets of the four proportions of the four-fold table. The results are given in Table 4. Of the 16,459 trials where inadequate results were obtained, the tests correctly detected the excessive error(s) on 15,169 trials (91%) while they were incorrect for 1,290 trials (9%). Thus, we may conclude that while the tests are not perfect, and could probably be considerably improved, they are clearly better than no tests at all which is the usual situation. Also, it is seen that of 416 trials when all parameter estimates are accurate, over half were incorrectly rejected. Thus, it is evident that these tests may only be useful for behavior ratings closely approximated by the model.

It should be noted that the parameter values used for the Monte Carlo study do not result from any kind of random sampling of values from the real world of course. Thus the numbers in Table 4 do not permit us to estimate, say, the probability that the estimations of the parameters P, p_b, p_n are of certain degree of accuracy when the method is applied to data of real phenomena. That would not be possible with Monte Carlo study.

The Monte Carlo study did not include the effects of random sampling. Thus, the results can only be applied to large enough samples.

Results

Several of the original behaviors were found to have base-rates too low for analysis of rater agreement. Some of these were collapsed, in an "armchair" fashion, into single "negative aggressive" (NA) category (such as DS, MI, and PN); similarly, others were put into a "miscellaneous aggressive" (MA) category (such as YE, VA and WH). Ten behaviors of the original 31 had sufficiently high base-rates so as to not require further collapsing. The joint agreement bivariate frequency distribution for the resulting twelve behavior categories is given in Table 5.

The parameter estimates of p_b, p_n, P for each of the twelve behaviors are given in Table 6. We see that according to the model the raters are extremely accurate in the detection of both the presence and the absence of each of the behaviors. Also, we hasten to add that all of the above consistency tests were passed for each behavior with the exceptions that WK, IN, and NO failed the requirements that $\phi < .50$ and $P < .15$, which was simply due to the relatively high base-rates.

Since the values of $\hat{p}_b$ and $\hat{p}_n$ were higher than anticipated another Monte Carlo study of the model was performed with $p_b = .90, .91, .92, .93, .94, .95$ and $p_n = .05, .06, .07, .08, .09, .10$, and with the other parameters varied as before. The results of this second study are given in Table 7. We see that for these values, passage of the tests did not occur even

once where the estimates were too erroneous.

Now we have strong evidence that the detection accuracy of presence and absence of behavior is extremely high and that the idealized model is a close approximation to the actual situation. Does this mean that the agreement rate for the presence of the behavior should be high in each case?

Above we considered the conditional probability

$$p_b = \Pr(\text{rating is } + \mid \text{behavior is present}).$$

Likewise, we can consider the conditional probability which is the inverse of p_b, namely,

$$V = \Pr(\text{behavior is present} \mid \text{rating is } +)$$

which we will call the validity of the + rating. This is a model-based parameter and therefore, it has meaning to the extent the model has verisimilitude. (We check the verisimilitude by the use of consistency tests.) It follows that

$$V = \frac{p_b P}{p^+}, \tag{14}$$

and Table 8 gives these validity values for each behavior.

The proportion agreement coefficient is

$$\Pr(+ \mid +) = \frac{2p^{++}}{p^{-+} + p^{+-} + 2p^{++}}.$$

This coefficient is simply the proportion of all the + ratings made by either rater which have a concordant mate or associated rating.

When $p^{-+} = p^{+-}$, as the model assumes, then

$$Pr(+|+) = \frac{p^{++}}{p^{+}} .$$

For $p_n \doteq 0$, and $p_b \doteq 1$, we see that

$$V \doteq Pr(+|+).$$

That the approximation is correct for the present data is illustrated in Table 8. The two instances where it does not hold (NA and PP) are the two cases where the sample size was small enough to cause sufficient error in the estimates of P, p_b, and p_n. Thus, we see that for the present data the two coefficients $Pr(+|+)$ and V are of about the same magnitude. For the present study they can thus be thought of as interchangeable.

Discussion

This method of analysis allows us to see that the behaviors with the lowest validities or the lowest agreement rates are those with the lowest base-rates. The ability of the raters to accurately detect either the presence or absence of any of the twelve behaviors is extremely high (the lowest value is .974) and evidently plays little role in the determination of the agreement rate or validity. We can see analytically that often the sole culprit that causes low validity is the base-rate P. Since

$$V = \frac{p_b P}{p_b P + p_n Q} ,$$

we have

$$\frac{dV}{dP} = \frac{p_b(Pp_b + Qp_n) - Pp_b(p_b - p_n)}{(p_bP + p_nQ)^2}$$

$$= \frac{p_np_b}{(p^+)^2},$$

which is always positive. Thus, if we can increase P, while p_n and p_b are held constant (as they clearly are for the twelve behaviors), then V will increase. In a nutshell, if we wish to have the validity of a low base-rate behavior be as high as that of a high base-rate behavior we must have much greater accuracy in the detection of the presence and/or absence of the low base-rate behavior. Illustrative values of V as a function of P are shown in Table 9. We see that if we require that V be at least .8, for example, and if $p_b = .99$ and $p_n = .01$, then P must be at least .04, and if $p_b = .95$ and $p_n = .05$ then P must be over .16.

The base-rate can be increased by combining similar behavior categories if the assumptions of the model are met for the new behavior category. Also, the setting in which the behaviors are rated may be changed to obtain a sufficiently high base-rate. On the other hand, it would generally be hazardous to simply encourage that more + ratings be recorded as this could simply increase p_n so that the validity V is decreased.

If we do wish to increase the validity V by improving the p_b and p_n values, it is more effective to reduce the p_n

value. We see this by considering the two derivatives

$$\frac{dV}{dp_b} = \frac{PQp_n}{p^{+2}} = Kp_n$$

and

$$\frac{dV}{dp_n} = \frac{-PQp_b}{p^{+2}} = -Kp_b,$$

$$\text{where } K = \frac{PQ}{p^{+2}}.$$

Typically $p_b >> p_n$, and it follows that an increase in V will be relatively small for an increase in p_b as compared to that for a decrease in p_n.

In the particular application reported above we conclude that the low rater agreement, when it occurred (for MA, NA, HR, and AT), was mainly a result of the low base-rate and not so much a result of rater carelessness, inattentiveness, lack of behavior dichotomy, ambiguity of the behavior definition, and so on. Further attention to these latter sources of disagreements should lead to better agreement but what has been shown is that increasing the base-rate will be more effective.

It has been implicitly assumed above that it is ultimately desired to have a sufficiently high validity coefficient V and that is why we are concerned with the rater agreement. While this may usually be the case, it is not always so. It may be, for example, that we only wish to know if the raters' perceptions of the behavior are similar. If it should happen,

for example, that the behavior under question has a base-rate of only .001 and that we are confident that $p_b \leqq .99$ and $p_n \geqq .01$, then from Table 8 we see that we should be delighted with a value of the validity coefficient V of about .09 if the consistency tests are passed!

REFERENCE NOTE

1. Meehl, P. E. Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion (Rep. PR-65-2). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1965.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. Advances in Clinical Assessment, Vol. 3. San Francisco: Jossey-Bass Publishers, 1975.

TABLE 1

THE RATES OF AGREEMENT AND DISAGREEMENT FOR A HYPOTHETICAL EXAMPLE

		Rater 2		
		+	-	
Rater 1	+	.05	.09	.14
	-	.09	.77	.86
		.14	.86	

TABLE 2

BEHAVIOR CODING SYSTEM

AP (Approval): A person gives clear gestural or verbal approval to another individual. Must include some clear indication of positive interest or involvement.

CM (Command): This category is used when an immediate and clearly-stated request or command is made to another person.

CN (Command Negative): A command which is very different in attitude from a reasonable command or request (CM).
(1) Immediate compliance is demanded.
(2) Aversive consequences are threatened if compliance is not immediate.
(3) A kind of sarcasm or humiliation directed to the receiver.

CO (Compliance): A person immediately does what is asked of him.

CR (Cry): Whenever a person cries.

DI (Disapproval): The person gives verbal or gestural disapproval of another person's behavior or characteristics.

DP (Dependency): When person is requesting assistance in doing a task that he is capable of doing himself, and it is an imposition on the other person to fulfill the request.

DS (Destructiveness): The person destroys, damages, or attempts to damage any (non-human) object; the damage need not occur, but the potential for damage must exist. This may also include an attack on an object, such as stamping on the floor or pounding on a table.

ED (Educate): Instruction, illustration, advice and demonstration.

ER (Error): Any unsuccessful attempt to perform assigned task or life skill. Includes breakage, spilling, and dropping materials.

LA (Laugh): A person laughs or smiles in a non-humiliating way.

LS (Life Skills): Performance of any target behavior selected by staff as needing specific encouragement and reinforcement; will vary from person to person.

MI (Misappropriation): Using, handling, or taking an object belonging to someone else (without permission), or against rules.

NC (Noncompliance): When a person does not do what is requested of him. Must be immediately preceded by a command.

TABLE 2, continued

NE (Negativism): A statement in which the verbal message is neutral, but which is delivered in a tone of voice that conveys an attitude of, "Don't bug; don't bother me."

NR (No Response): When a person does not respond to another person. Applicable when a behavior does not require a response, or when behavior is directed at another person, but the person to whom the behavior is directed fails to perceive the behavior (or ignoring).

PL (Play): A person is playing either alone or with other persons.

PN (Physical Negative): A subject physically attacks or attempts to attack another person with sufficient intensity to potentially inflict pain.

PP (Physical Positive): Positive touching, a pat on the back or a hug.

RR (Receives): Receives any item of potential value.

TE (Tease): Teasing another person in such a way that the other person is likely to show displeasure and disapproval or when the person being teased is trying to do some other behavior, but is unable to because of the teasing.

VA (Vanish): Absent without leave, wanders off, walks away from task situation.

WH (Whine): A person states something in a slurring, nasal, high-pitched, falsetto voice.

WK (Work): A person is working, either alone or with other people.
(1) Assigned task.
(2) The behavior is necessary for a child to perform in order to learn behavior which will help him assume an adult role.

YE (Yell): The person shouts, yells, or talks loudly.

AT (Attention): When one person listens to or looks at another person, and the categories AP or DI are not appropriate.

NO (Normative): A person is behaving in an appropriate fashion and no other code is applicable.

OO (Other Objectionables): Other objectionable behavior not classified elsewhere.

SS (Self-Stimulation): Repetitive behaviors which the individual does to himself and cannot be coded by any other codes (includes talking to self).

TA (Talking): This code is used if none of the other verbal codes are applicable (questions not falling under any first order category are included).

TH (Touch): When the subject touches other people or hands an object to another person.

TABLE 3

THE PROPORTIONS OF THE FOUR-FOLD TABLE

		Rater 2		
		+	-	
Rater 1	+	p^{++}	p^{+-}	p_1^{+}
	-	p^{-+}	p^{--}	p_1^{-}
		p_2^{+}	p_2^{-}	

where $p^{++} + p^{+-} + p^{-+} + p^{--} = 1$

$$p_1^{+} + p_1^{-} = 1$$

$$p_2^{+} + p_2^{-} = 1$$

$$p_1^{+} = p^{++} + p^{+-}$$

$$p_1^{-} = p^{-+} + p^{--}$$

$$p_2^{+} = p^{++} + p^{-+}$$

$$p_2^{-} = p^{+-} + p^{--}$$

TABLE 4

RESULTS OF THE FIRST MONTE CARLO STUDY

	Consistency Test Outcome		
	Accept	Reject	Total
Accurate Estimates[a]	166	250	416
Inaccurate Estimates	1,290	15,169	16,459
Total	1,456	15,419	16,875

Note. The frequencies refer to the number of artificial data trials.

[a]All of the parameter estimate errors are less than or equal to .10.

TABLE 5

JOINT AGREEMENT BIVARIATE FREQUENCY DISTRIBUTION

Rater 2	Rater 1 MA	NA	LA	PL	PP	WK	HR	IN	AT	NO	SS	TA	Total
MA	102	1	1	3	0	6	8	4	1	21	7	24	178
NA	5	35	3	2	2	1	3	3	0	13	0	6	73
LA	5	3	410	0	0	31	5	62	6	36	18	14	590
PL	2	2	2	118	0	20	3	15	0	32	2	5	201
PP	6	7	1	0	15	1	0	10	0	0	0	2	42
WK	14	0	14	4	0	7234	17	193	62	82	83	31	7734
HR	0	1	4	2	0	8	101	15	1	11	5	0	148
IN	7	0	61	0	0	238	6	4881	25	221	195	38	5672
AT	2	1	9	1	2	78	2	46	265	32	15	33	486
NO	29	7	35	5	2	84	16	137	10	3888	138	80	4431
SS	2	0	20	1	1	85	10	265	5	194	3296	34	3913
TA	13	8	21	13	2	43	3	50	24	70	58	886	1191
Total	187	65	581	149	24	7829	174	5681	399	4600	3817	1153	24659

TABLE 6

PARAMETER ESTIMATES FOR EACH OF THE TWELVE BEHAVIORS

Behavior	$\hat{P}$	$\hat{p}_b$	$\hat{p}_n$
MA	.004	.997	.997
NA	.001	.998	.998
LA	.017	.993	.993
PL	.005	.998	.998
PP	.001	.999	.999
WK	.306	.978	.977
HR	.004	.997	.997
IN	.211	.967	.967
AT	.011	.993	.993
NO	.166	.974	.974
SS	.140	.977	.976
TA	.037	.988	.988

TABLE 7

RESULTS OF THE SECOND MONTE CARLO STUDY

	Consistency Test Outcome		
	Accept	Reject	Total
Accurate Estimates[a]	2,889	11,180	14,069
Inaccurate Estimates	0	2,806	2,806
Total	2,889	13,986	16,875

Note. The frequencies refer to the number of artificial data trials.

[a]All of the parameter estimate errors are less than or equal to .10.

TABLE 8

RATER AGREEMENT AND VALIDITY COEFFICIENT VALUES

Behavior	$\hat{P}$	p^+	$Pr(+\mid+)$	V
MA	.004	.007	.559	.539
NA	.001	.003	.507	.356
LA	.017	.024	.700	.711
PL	.005	.007	.674	.703
PP	.001	.001	.454	.746
WK	.306	.316	.930	.948
HR	.004	.007	.627	.611
IN	.211	.216	.860	.946
AT	.011	.018	.600	.609
NO	.166	.183	.861	.883
SS	.140	.157	.853	.873
TA	.037	.048	.756	.769

p^+: proportion of all ratings which are +

$Pr(+\mid+)$: probability that rating of one rater will be + if other rating is +

V: model-based validity coefficient

TABLE 9

THE VALIDITY V AS A FUNCTION OF THE BASE-RATE P FOR TWO DIFFERENT VALUES OF p_b AND p_n

P	V	
	p_b=.99, p_n=.01	p_b=.95, p_n=.05
.001	.090	.019
.005	.332	.087
.01	.500	.161
.02	.669	.279
.04	.805	.442
.08	.896	.623
.16	.950	.783