Open access • Journal Article • DOI:10.1177/014662169201600109

Estimating individual rater reliabilities — Source link

John E. Overall, Kevin N. Magee

Institutions: University of Texas at Austin

Published on: 01 Mar 1992 - Applied Psychological Measurement (SAGE Publications)

Topics: Inter-rater reliability

Estimating Individual Rater Reliabilities

John E. Overall and Kevin N. Magee
University of Texas Medical School

Rating scales have no inherent reliability that is independent of the observers who use them. The often reported interrater reliability is an average of perhaps quite different individual rater reliabilities. It is possible to separate out the individual rater reliabilities given a number of independent raters who observe the same sample of ratees. Under certain assumptions, an external measure can replace one of the raters, and individual reliabilities of two independent raters can be estimated. In a somewhat similar fashion, estimates of treatment effects present in ratings by two independent raters can provide the external frame of reference against which differences in their individual reliabilities can be evaluated. Models for estimating individual rater reliabilities are provided for use in selecting, evaluating, and training participants in clinical research.
Index terms: attenuation, correlation, individual raters, interrater reliability, multiple raters, rater reliability, rating scales, reliability of ratings, significance.

Rating scales provide essential measurements for much clinical research. References to the reliability of a rating instrument are common; however, clinical rating scales do not have inherent reliability that is independent of the skill of the observers who use them. The fact that raters differ in the accuracy of their judgments, and in the consistency with which those judgments are recorded, is a major problem for clinical research. This paper describes several models that provide estimates of the reliability of a rating scale used by a particular individual rater. The use of such models may be helpful in selecting and training raters for participation in clinical research.

Concern with interrater consistency has generally focused on the reliability of the rating instrument or procedure. Estimates of the reliability of mean ratings calculated across raters and estimates of the average reliability of the individual raters using an instrument have been provided (Armstrong, 1981). Haggard (1958) and Winer (1962) were among the first to introduce psychologists to the intraclass correlation coefficient as a measure of average interrater reliability, and Shrout and Fleiss (1979) have elaborated the statistical models underlying the intraclass correlation coefficient. Cronbach, Gleser, Nanda, and Rajaratnam (1972) extended the components of variance approach to complex experimental designs in which variability among different raters is one factor influencing "generalizability." However, none of those authors discuss the problem of separating differing individual rater reliability as such. More recent authors have compared other approaches to estimating overall consistency among several raters as distinct from the specific reliabilities of individual raters (James, Demaree, & Wolf, 1984; Jones, Johnson, Butler, & Main, 1983; Towstopiat, 1984).

Coefficient alpha (Cronbach, 1951) is another frequently used estimate of the reliability of a composite test or the average reliability of subtests entering into the composite (Conn & Ramanaiah, 1990; Dolan, Lacey, & Evans, 1990). Coefficient alpha can be calculated (for standardized ratings) from the correlation matrix relating ratings made by several raters; in that regard, it is more logically compared with the models for individual rater reliabilities discussed here than is the intraclass coefficient obtained from an analysis of variance model. This comparison emphasizes further the difference between average rater reliability and individual rater reliability. It also suggests that the models discussed

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 77-85

0146-6216/92/010077-09$1.70

here potentially may be used for examining differential subtest reliabilities in cases where coefficient alpha may be used to obtain an internal consistency estimate for reliability of the composite score.

Only a few recent authors have specifically addressed the problem of concern in this article. Dillon and Mulani (1984) considered a simple observational setting in which several raters provide categorical judgments pertaining to each of several objects or individuals. The categorical data are then subjected to latent class analysis to provide estimates of the probability of each possible response pattern across the several raters. Although the "error rates" estimated in that manner are not reliability coefficients in the usual sense, they can be used to order judges according to their agreement with a consensus defined by parameters of the latent class model.

Van den Bergh and Eiting (1989) used another complex mathematical approach to evaluate individual rater reliabilities. They used LISREL (Jöreskog & Sorbom, 1988) to fit models that alternatively assume multiple quantitative ratings to be congeneric, tau-equivalent, or parallel. In spite of the complexity of this general structural equations approach, they provided results for the three-rater congeneric case that were equivalent to the disattenuation models described below.

By using simplifying assumptions common to classical test theory, formulas are provided below for estimating individual rater reliabilities in terms of the simple pairwise correlations among the ratings. As will be emphasized, several estimates of individual rater reliability that are proposed were anticipated by Lord and Novick (1968) in application to multiple parallel test forms. The intent of this paper was to emphasize that individual raters may have different reliabilities that may be important to evaluate in clinical research settings. Reliability here concerns the accuracy of discriminating true differences among individuals or objects.

The models considered here assume that multiple raters, like parallel forms of a test, are attempting to measure the same trait, symptom, or behavior, and that the ratings would correlate perfectly if the raters were perfectly reliable. Lack of correlation is considered measurement error. Jöreskog (1971) has used the term "congeneric" to indicate ratings that measure the same trait except for errors of measurement. In some cases, an instrument of a different sort (e.g., a psychological test) may substitute for another rater in the derivation of individual rater reliability. In a somewhat similar fashion, the ability to discriminate between fixed treatment conditions can serve as the basis for assessing individual rater reliability. Although correlation with a related measure or discrimination of treatment effects is usually considered evidence of validity, such external frames of reference can provide a basis for distinguishing between reliable and less reliable ratings. Artificial data were used to illustrate the simple calculations in the following cases.

Models for Rater Reliability

Disattenuation Model

Assume that ratings by three raters would correlate perfectly if it were not for measurement error. This assumption implies that the disattenuated interrater correlations should be unity. Disattenuation results when the observed correlation between each pair of raters is divided by the geometric mean of their individual reliabilities (Lord & Novick, 1968, p. 70). For three raters, this defines the following pairwise equations:

$$r_{12}/\sqrt{r_{11}}\sqrt{r_{22}} = 1.0$$

$$r_{13}/\sqrt{r_{11}}\sqrt{r_{33}} = 1.0$$

$$r_{23}/\sqrt{r_{22}}\sqrt{r_{33}} = 1.0 \quad , \tag{1}$$

where r_{12}, r_{13}, and r_{23} are the observed correlations between ratings by the three raters, and r_{11}, r_{22}, and r_{33} are their individual reliabilities.

Simultaneous solution yields the following formulas for estimating individual rater reliabilities:

$$r_{11} = r_{12}\frac{r_{13}}{r_{23}}$$

$$r_{22} = r_{12}\frac{r_{23}}{r_{13}}$$

$$r_{33} = r_{23}\frac{r_{13}}{r_{12}} \quad . \tag{2}$$

Lord and Novick (1968, pp. 216–218) derived formulas that are essentially equivalent to these and to other results based on disattenuation. Their derivation began with a simple measurement model in which observed scores are combinations of true and error components. Their solution involved the covariances (rather than correlations) of the observed scores. Lord and Novick's covariance formulas provided estimates of the "true variance" for individual tests, from which individual test reliability was calculated as the ratio of true variance to total observed variance. In spite of the fact that Lord and Novick used a different point of departure and were not concerned specifically with the reliabilities of individual raters, their results clearly anticipate those presented here. As noted, Van den Bergh and Eiting (1989) employed a general structural equations approach to evaluate individual rater reliabilities; however, for the specific case of three raters, the solution in their Equations 17 through 19 is the same as the disattenuation solution presented here. Again, although the present derivation is based on the disattenuation of measurements of the same trait by several observers, it is the separation of individual rater reliabilities from instrument reliability averaged across raters that is emphasized as a problem of practical importance for clinical research.

The relationship between interrater correlations and measurement reliabilities defined in Equation 1 is important in classical measurement theory (Lord & Novick, 1968, p. 70), but it does not account for the sampling error in observed correlations. Although the relationship $r_{12} / (r_{11})^{1/2}(r_{22})^{1/2}$ is discussed in numerous texts to disattenuate the observed correlation between two measurements, the disattenuated correlation may exceed 1.0 if estimates of reliability are inaccurate. Reversing the problem, using Equation 2 to obtain estimates of individual rater reliabilities is subject to sampling error that is present in the observed correlations among the ratings. This tendency varies inversely with sample size. An example is used to illustrate calculation of the individual rater reliabilities estimated by this and other models, and a test of significance for differences among the individual rater reliabilities is discussed below.

Table 1 presents intercorrelations among ratings of "thinking disturbance" recorded on the Brief Psychiatric Rating Scale (Overall & Gotham, 1962) by three raters who independently interviewed and rated each of 25 schizophrenic patients. Each patient also completed the short form of the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway & McKinley, 1982). Correlations of the Sc (Scale 8) scores with the clinical ratings are also presented in Table 1 for later reference. The individual reliabilities for the three raters estimated from the disattenuation model are as follows:

$$\text{Rater 1: } r_{11} = .88\frac{.67}{.64} = .92$$

$$\text{Rater 2: } r_{22} = .88\frac{.64}{.67} = .84$$

$$\text{Rater 3: } r_{33} = .64\frac{.67}{.88} = .49 \quad . \tag{3}$$

Table 1
Intercorrelations Among Three Raters and an External Measure

Rater	Rater 1	Rater 2	Rater 3
2	.88		
3	.67	.64	
MMPI	.52	.49	.38

Common Factor Model

Again assume that three (or more) raters all rate the same trait or behavior for a number of ratees and that these ratings would correlate perfectly if it were not for measurement error. According to the common factor model, observed correlations among ratings made by different raters arise because they all reflect the same underlying unitary source of variance, namely true individual differences on the trait or behavior being rated.

It is postulated that a single common factor should account for all the observed intercorrelations among ratings by the different raters if they are indeed measuring the same unitary trait or behavior, except for measurement error (Lord & Novick, 1968, p. 536). The so-called "fundamental theorem" of factor analysis requires that cross-products of the factor loading coefficients should reproduce the observed intercorrelations (Harman, 1960, p. 35). Specifically, the first principal component (principal axes factor) of a matrix of intercorrelations with communality estimates in its principal diagonal provides a unique solution such that cross-products of the factor loading coefficients approximate maximally the observed correlations. No other single set of coefficients will as closely reproduce all the observed correlations. Given intercorrelations among ratings provided by three raters, the single-factor model specifies the following relationship:

$$r_{12} = f^{(1)}f^{(2)}$$
$$r_{13} = f^{(1)}f^{(3)}$$
$$r_{23} = f^{(2)}f^{(3)} \quad , \tag{4}$$

where $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ are loadings on the first common factor of the matrix of intercorrelations among ratings by the three raters.

Note the similarity between Equations 1 and 4. Dividing both sides of Equation 4 by $f^{(i)}f^{(j)}$ suggests that $f^{(i)}f^{(j)} = (r_{ii})^{1/2}(r_{jj})^{1/2}$.

A principal axes factor analysis of the matrix of intercorrelations among the three raters (extracted from Table 1) produced the following vector of factor loadings:
{.958 .918 .698}.

The estimates of individual rater reliabilities derived from the common factor model are as follows:
$r_{11} = (.958)(.958) = .93$
$r_{22} = (.918)(.918) = .84$
$r_{33} = (.698)(.698) = .49$

Under the single-factor model, the individual reliability estimates are also communality estimates. Thus, if a computer program for factor analysis is used, the individual rater reliability estimates are provided directly as communality values. Lord and Novick (1968, p. 536) also discussed the relation of communality and reliability in cases where tests may contain specific variance that is not random measurement error. In that case, communality estimates can be taken as a lower bound for reliability.

The difference between the attenuation and factor models is that the factor model does not

require that a single factor must perfectly reproduce the complete set of observed correlations between all pairs of raters, which may not be possible when there are several. The factor model is also more convenient when ratings made by several raters are the basis for estimating their individual reliabilities.

External Criterion Model

In this model, the measurement of a related, although not necessarily identical, trait separates the individual rater reliabilities. Because the external measure (e.g., a psychological test) may not correlate perfectly with the ratings even after the effects of random measurement errors are removed, the correlation with the external measure must be "disattenuated" for "specific variance" that does not depend on the trait or behavior that is being rated.

Let two raters independently rate the level of a specified trait or behavior for a number of individuals, and let a related trait be measured by a psychological test. This model begins with the same disattenuation model, but adjusts the observed correlations between the test and the two sets of ratings for the specific variance in the test scores. A critical assumption of this model is that the proportion of the total variance of the test scores that is related to the trait or behavior being rated is unaffected by the differential reliabilities with which the ratings are recorded by the raters. The common component, designated c_{33}, is equal to the total (standardized) variance of the test scores minus the proportions of that total variance due to both specific variance and measurement error, and $1 - c_{33}$ is the "uniqueness" for which the observed test/rating correlations must be disattenuated if they are to equal 1.0. Thus,

$$\begin{aligned} r_{12}/\sqrt{r_{11}}\sqrt{r_{22}} &= 1.0 \\ r_{13}/\sqrt{r_{11}}\sqrt{c_{33}} &= 1.0 \\ r_{23}/\sqrt{r_{22}}\sqrt{c_{33}} &= 1.0 \quad , \end{aligned} \tag{5}$$

where r_{12} is the observed correlation between Raters 1 and 2, and r_{13} and r_{23} are those raters' observed correlations with the external criterion. Again, it is assumed that the ratings by the two raters would correlate perfectly if it were not for random measurement error, but it is not assumed that the external criterion would correlate perfectly with the ratings if it were not for random measurement error.

Simultaneous solution of this set of equations yields formulas for estimating individual rater reliabilities that appear identical to those provided by the three-rater disattenuation model (Equation 2, eliminating Rater 3). The individual rater reliabilities are equal to the observed correlation between the raters adjusted up or down according to whether the particular rater correlates with the external criterion to a greater or a lesser degree than does the other rater. At first, it may seem counterintuitive that the same equations estimate individual rater reliabilities whether or not the third set of scores is assumed to measure exactly the same thing except for random measurement error. However, it is the ratio of the correlations with the external criterion that is important, not the absolute magnitudes of those correlations.

Lord and Novick (1968, pp. 216–218) provided an equivalent solution in terms of covariances among scores that are not truly parallel. This model suggests that individual rater reliabilities can be scaled against external criteria that have only modest correlations with the ratings. Although that is true, ratios between observed correlations tend to be unstable when the absolute values are low, because the low correlations may be due to chance. Thus, the external criterion selected should be a valid measure of the trait or behavior that is also being rated.

The MMPI Sc correlations from Table 1 can be used to illustrate calculation of individual rater reliabilities for Raters 1 and 2 using an external criterion:

$$\text{Rater 1: } r_{11} = .88\frac{.52}{.49} = .93$$
$$\text{Rater 2: } r_{22} = .88\frac{.49}{.52} = .83 \quad . \tag{6}$$

Suppose that the individual rater reliabilities of Raters 1 and 3 were of interest, rather than the individual reliabilities of Raters 1 and 2. This provides a way of examining the stability of estimates for Rater 1 derived from correlation with a different rater and an external criterion:

$$\text{Rater 1: } r_{11} = .67\frac{.52}{.38} = .92$$
$$\text{Rater 3: } r_{33} = .67\frac{.38}{.52} = .49 \quad . \tag{7}$$

If only the MMPI Sc number-correct scores and Raters 2 and 3 are available, the following estimates of their individual reliabilities would be produced by the correlation coefficients in Table 1:

$$\text{Rater 2: } r_{22} = .64\frac{.49}{.38} = .83$$
$$\text{Rater 3: } r_{33} = .64\frac{.38}{.49} = .50 \quad . \tag{8}$$

Treatment Effects Model

Magee and Overall (in press) have provided formulas for estimating the individual reliabilities of two raters from differences in the magnitudes of components of variance for treatment effects and residual error in an experiment in which two raters independently evaluate each ratee in two treatment groups. That model is not discussed in detail here because summary statistics other than simple bivariate correlations (e.g., Table 1) are required. However, treatment group membership can be arbitrarily coded [1,–1], and the resulting dummy variable can be treated as an external criterion for correlation with ratings of treatment outcome by two raters. Formulas for estimating individual rater reliabilities from correlations with an external criterion can then be used as an alternative to the components of variance calculation.

Regression Model

The regression model assumes that the ratings provided by a rater should be predicted perfectly from ratings provided by other raters if it were not for measurement error. A limitation is that such perfect prediction would require the other raters to be perfectly reliable, which they are not. Considering this limitation, the multiple R^2 relating ratings made by one rater to the ratings provided by other raters and/or external measures is an estimate of the individual rater's reliability. The multiple R^2 is attenuated to some degree by the unreliability of the "independent variables" in the regression equation; hence, the individual rater reliabilities estimated in this way tend to underestimate the actual values. Nevertheless, the regression estimates may be useful in distinguishing more reliable raters from less reliable raters. It was noted above that estimates of individual reliability derived from the common factor model are also communality estimates in factor analysis. Now note that the multiple R^2 estimates relating each variable to all other variables are the initial communality values used by

many computer programs for factor analysis (e.g., SPSS, SAS, BMDP).

The regression model can be used to estimate individual reliabilities for three or more raters using a combination of ratings and/or external criterion measures. To illustrate use of the regression model, the REGRESSION procedure from SPSS (Norušis, 1986) was used to calculate the following standardized multiple regression equations and multiple R^2 values by entering the 4 × 4 correlation matrix from Table 1. The results were

Y_1 = (.722)(Rater 2) + (.169)(Rater 3) + (.102)(MMPI) R_1^2 = .80,
Y_2 = (.799)(Rater 1) + (.089)(Rater 3) + (.041)(MMPI) R_2^2 = .78, and
Y_3 = (.460)(Rater 1) + (.219)(Rater 2) + (.034)(MMPI) R_3^2 = .46.

In this case, each rater's ratings were regressed on a combination of the other two raters and the MMPI Sc scores. This illustrates the use of a combination of ratings by other raters and an external test for estimating rater reliabilities.

It is apparent from these results that the multiple regression model systematically underestimated the individual rater reliabilities. Cronbach et al. (1972, pp. 312–314, 351) considered the reduced variability of regression estimated scores to be an advantage, but monte carlo studies by the present authors have confirmed that the reliability estimates obtained in this way do, in fact, underestimate the actual reliability built into simulated data. This problem increases when only a few raters are used, rather than several. Thus, the regression model is more appropriately used to estimate individual reliabilities when there are several raters who rate the same ratees. However, using a common factor or disattenuation model is preferable to the regression model for accuracy in estimating individual rater reliabilities from simple correlational data.

Individual Rater Reliabilities and Coefficient Alpha

As noted above, coefficient alpha can be used to obtain an estimate of the average reliability for several raters. Coefficient alpha for the three raters was calculated from the intercorrelations among their ratings that are shown in Table 1. The reliability for a composite score calculated as the sum of the (standardized) ratings by the three raters was α_3 = .89. The average reliability for the individual raters that was obtained by a reverse Spearman-Brown calculation for a test of one-third length was α_1 = .73. The mean of the three individual reliabilities calculated for the disattenuation model (Equation 2) was r_{tt} = .75. This suggests, but does not prove for the general case, that the disattenuation calculations closely estimate the individual reliabilities for three raters whose mean reliability is defined by coefficient alpha. Coefficient alpha provides a lower bound estimate for reliability that depends on measurements being parallel. The disattenuation model assumes that the raters are rating the same thing except for random measurement error. If that is not true, both coefficient alpha and the individual rater reliabilities will underestimate the actual reliability of the ratings.

Significant Differences Among Individual Rater Reliabilities

It was noted above that the estimates of individual rater reliabilities are subject to sampling error. An important question to ask of observed estimates of individual rater reliabilities is whether the difference among them is greater than might be expected to occur by chance if the different raters are actually equally reliable. Because the estimates of individual rater reliabilities depend solely on observed intercorrelations among their ratings, the test of significance can be applied directly to the observed correlations. If the raters are rating the same variable except for measurement error, then a significant difference in the correlations among their ratings is evidence of true differences in reliability.

A test of significance for the difference between two nonindependent correlation coefficients that

was developed by Hotelling (1940) is proposed as a criterion for evaluating the difference between the largest and smallest of three observed pairwise correlations in the case of three raters (e.g., the disattenuation or common factor model). Let the largest and smallest observed correlations be r_{13} and r_{23}, in which case the estimated reliabilities of Raters 1 and 2 will be the most disparate. Hotelling's formula for a t statistic with $n - 3$ degrees of freedom (df) is

$$t = (r_{13} - r_{23})\left[\frac{(n-3)(1+r_{12})}{2(1 - r_{13}^2 - r_{23}^2 - r_{12}^2 + 2r_{13}r_{23}r_{12})}\right]^{1/2} . \quad (9)$$

Consider the question of statistical significance of differences in reliabilities of ratings by the three raters that were presented in Table 1. For consistency with the notation of Equation 9, the largest correlation is designated as $r_{13} = .88$, the smallest as $r_{23} = .64$, and the remaining intermediate value as $r_{12} = .67$. Substituting into Equation 9, $t = 2.91$ with 22 df. The two-tailed p value associated with this t statistic is $p < .01$.

Recommendations and Conclusions

Rating scales do not have inherent reliability that is independent of the skill and training of the raters. The frequently reported "interrater reliability," the intraclass correlation, and coefficient alpha are averages of the individual reliabilities of the two or more raters from whom the reliability data derive. Several simple models for estimating individual rater reliabilities from simple bivariate correlations among their ratings were proposed. Because such estimates are subject to sampling variability, a reasonable N is required for confidence that the apparent differences are real, and a test of significance applied to differences among the pairwise correlations can be a safeguard against erroneous conclusions.

References

Armstrong, G. D. (1981). The intraclass correlation as a measure of interrater reliability of subjective judgments. *Nursing Research, 30,* 314–315.

Conn, S. R., & Ramanaiah, N. V. (1990). Factor structure of the Comrey Personality Scales, the Personality Research Form E, and the five-factor model. *Psychological Reports, 67,* 627–632.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19,* 438–458.

Dolan, B., Lacey, J. H., & Evans, C. (1990). Eating behavior and attitudes to weight and shape in British women from three ethnic groups. *British Journal of Psychiatry, 157,* 523–528.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* New York: Dryden Press.

Harman, H. H. (1960). *Modern factor analysis.* Chicago: The University of Chicago Press.

Hathaway, S. R., & McKinley, J. C. (1982). *Minnesota Multiphasic Personality Inventory: Group form test booklet.* Minneapolis: University of Minnesota Press.

Hotelling, H. (1940). The selection of variates for use in prediction, with some general comments on nuisance parameters. *Annals of Mathematical Statistics, II,* 271–283.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85–98.

Jones, A. P., Johnson, L. A., Butler, M. C., & Main, D. S. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. *Academy of Management Journal, 26,* 507–519.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–132.

Jöreskog, K. G., & Sorbom, D. (1988). *LISREL VII: A guide to program applications* [Computer program and manual]. Chicago: SPSS, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Magee, K., & Overall, J. E. (in press). Improved formulae for estimating individual rater reliabilities from observed treatment effects. *Educational and Psychological Measurement.*

Norušis, M. T. (1986). *SPSS/PC+: SPSS for the IBM PS/XT/AT.* Chicago: SPSS Inc.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799–812.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Towstopiat, O. (1984). A review of reliability procedures for measuring observer agreement. *Contemporary Educational Psychology, 9,* 333–352.

Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement, 26,* 29–40.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Acknowledgments

This work was supported in part by grant 5R01 MH 32457-11 NIMH.

Author's Address

Send requests for reprints or further information to John E. Overall, Department of Psychiatry, University of Texas Medical School, P.O. Box 20708, Houston TX 77225, U.S.A.