PSYCHOMETRIKA—VOL. 26, NO. 2
JUNE, 1961

ANALYSIS OF UNREPLICATED THREE-WAY CLASSIFICATIONS, WITH APPLICATIONS TO RATER BIAS AND TRAIT INDEPENDENCE*

JULIAN C. STANLEY

UNIVERSITY OF WISCONSIN

The seven analysis-of-variance mean squares for an unreplicated three-way classification may be written as linear combinations of a mean variance and three mean covariances. Formulas are presented for computing the mean variances and mean covariances from linear combinations of mean squares. The relevance of these formulas for assessing rater biases and trait independence is discussed, a numerical example is provided, and proposed extensions are briefly noted.

When repeated measurements of individuals are made over all levels of two experimental variables, three sources of covariance become possible. Consider the familiar situation where each individual is rated once by each rater on each trait, there being at least two individuals, two raters, and two traits. Covariation can occur within each rater across traits, within each trait across raters, and across both raters and traits.

These three sources of covariation are orthogonal. Empirically, it has been found that covariation within raters across traits, inflated by relative halo effect, usually exceeds covariation within traits across raters, the magnitude of which reflects independence of the traits. Covariation across both raters and traits constitutes a baseline against which the other two sources may be judged. It tends to be less than either of them.

In 1954, Guilford ([8], p. 281) showed that the various rater biases can be thought of appropriately in terms of analysis-of-variance mean squares involving raters: the mean squares for (i) raters, (ii) the interaction of raters with ratees, (iii) the interaction of raters with traits, and (iv) the second-order interaction of raters with both ratees and traits. Thus, there are four possible sources of rater bias, three of which (the main effect of raters and the two first-order interactions) may be evaluated in a given study and compensated for statistically, as will be shown in this article.

The ratee-rater-trait matrix is used above merely as an introductory illustration. Also, ratees define rows only for convenience of initial exposi-

*The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare. The assistance of Sister M. Jacinta Mann, S. C., at one stage of this investigation is gratefully acknowledged.

tion. Raters or traits might just as well define rows. In each instance, three sources of covariation can be isolated. Over the three orderings, each first-order interaction will be defined twice and the second-order interaction three times. In a generalized, complete three-classification matrix each interaction mean square may be shown to be a linear function of a mean within-column variance and three mean covariances among "levels" of the two factors defining columns.

Method of Analysis

Consider any matrix of real numbers X_{irt}, where $i = 1, 2, \cdots, I$; $r = 1, 2, \cdots, R$; and $t = 1, 2, \cdots, T$. Partition the total sum of squared deviations around the mean of these $I \times R \times T$ numbers into the usual seven sums of squares: three for main effects, three for first-order (two-factor) interactions, and one for second-order (three-factor) interaction.

The four mean squares (i.e., sums of squares divided by the appropriate number of degrees of freedom) involving i may be written

$$\text{(1)} \quad \text{MS}_i = A + (R-1)B + (T-1)C + (R-1)(T-1)D,$$

$$\text{(2)} \quad \text{MS}_{(i\times r)} = A - B + (T-1)C - (T-1)D,$$

$$\text{(3)} \quad \text{MS}_{(i\times t)} = A + (R-1)B - C - (R-1)D,$$

$$\text{(4)} \quad \text{MS}_{(i\times r\times t)} = A - B - C + D,$$

where

$$A = \overline{s_{rt}^2}, \; B = \overline{\text{cov}\,(X_{rt}, X_{rt'})}, \; C = \overline{\text{cov}\,(X_{rt}, X_{rt'})}, \; D = \overline{\text{cov}\,(X_{rt}, X_{r't'})}$$

with $r \neq r'$ and $t \neq t'$. Bars denote means. (For an outline of the way in which the formulas were obtained, see the Appendix at the end of this paper.)

If, for convenience, the i factor is considered to define rows of the matrix and the other two factors columns, A is the mean of the RT within-column variances of the form

$$s_{rt}^2 = \sum_{i=1}^{I} (X_{irt} - \bar{X}_{\cdot rt})^2/(I-1).$$

B is the mean of the $T[R(R-1)]$ covariances across the R raters within the T traits. C is the mean of the $R[T(T-1)]$ covariances across t's within r's. D is the mean of the remaining $RT(RT-1) - RT(R-1) - RT(T-1) = RT(R-1)(T-1)$ covariances, those across both r's and t's.

Formulas (1)–(4), independent linear equations in four unknowns, can be solved for the mean variance and the three mean covariances to secure the following formulas, where $MS_i = w$, $\text{MS}_{(i\times r)} = x$, $\text{MS}_{(i\times t)} = y$, and $\text{MS}_{(i\times r\times t)} = z$.

$$\text{(5)} \quad A = [w + (R-1)x + (T-1)y + (R-1)(T-1)z]/RT,$$

$$B = [w - x + (T-1)y - (T-1)z]/RT, \tag{6}$$

$$C = [w + (R-1)x - y - (R-1)z]/RT, \tag{7}$$

$$D = [w - x - y + z]/RT. \tag{8}$$

By treating factor r as defining rows and factors i and t as defining columns, one obtains expressions analogous to those of (1)–(4):

$$\mathrm{MS}_r = E + (I-1)F + (T-1)G + (I-1)(T-1)H, \tag{9}$$

$$\mathrm{MS}_{(r\times i)} = E - F + (T-1)G - (T-1)H, \tag{10}$$

$$\mathrm{MS}_{(r\times t)} = E + (I-1)F - G - (I-1)H, \tag{11}$$

$$\mathrm{MS}_{(r\times i\times t)} = E - F - G + H, \tag{12}$$

where

$$E = \overline{s^2_{it}}, \qquad F = \overline{\mathrm{cov}\,(X_{it}, X_{i't})},$$
$$G = \overline{\mathrm{cov}\,(X_{it}, X_{it'})}, \qquad H = \overline{\mathrm{cov}\,(X_{it}, X_{i't'})},$$

with $i \neq i'$ and $t \neq t'$.

Solving (9)–(12), one obtains the following formulas, where

$$\mathrm{MS}_r = u, \quad \mathrm{MS}_{(r\times i)} = \mathrm{MS}_{(i\times r)} = x, \quad \mathrm{MS}_{(r\times t)} = v,$$
$$\mathrm{MS}_{(r\times i\times t)} = \mathrm{MS}_{(i\times r\times t)} = z.$$

$$E = [u + (I-1)x + (T-1)v + (R-1)(T-1)z]/IT, \tag{13}$$

$$F = [(T-1)(v-z) + u - x]/IT, \tag{14}$$

$$G = [(I-1)(x-z) + u - v]/IT, \tag{15}$$

$$H = [(u - x - v + z)]/IT. \tag{16}$$

Finally, treating t as defining rows and i and r as defining columns,

$$\mathrm{MS}_t = J + (I-1)K + (R-1)L + (I-1)(R-1)M, \tag{17}$$

$$\mathrm{MS}_{(t\times i)} = J - K + (R-1)L - (R-1)M, \tag{18}$$

$$\mathrm{MS}_{(t\times r)} = J + (I-1)K - L - (I-1)M, \tag{19}$$

$$\mathrm{MS}_{(t\times i\times r)} = J - K - L + M, \tag{20}$$

where

$$J = \overline{s^2_{ir}}, \qquad K = \overline{\mathrm{cov}\,(X_{ir}, X_{i'r})},$$
$$L = \overline{\mathrm{cov}\,(X_{ir}, X_{ir'})}, \qquad M = \overline{\mathrm{cov}\,(X_{ir}, X_{i'r'})},$$

with $i \neq i'$ and $r \neq r'$. Solving (17)–(20) one obtains the following formulas,

where

$$\mathrm{MS}_{t} = q, \quad \mathrm{MS}_{(i\times i)} = y, \quad \mathrm{MS}_{(i\times r)} = v, \text{ and } \mathrm{MS}_{(i\times i\times r)} = z.$$

$$J = [q + (I - 1)y + (R - 1)v + (I - 1)(R - 1)z]/IR, \tag{21}$$

$$K = [(R - 1)(v - z) + q - y]/IR, \tag{22}$$

$$L = [(I - 1)(y - z) + q - v]/IR, \tag{23}$$

$$M = [(q - y - v + z)]/IR. \tag{24}$$

Note that each of the two-factor interactions is defined twice while the three-factor interaction is defined thrice. For example, by (2) and (10),

$$\begin{aligned}\mathrm{MS}_{(i\times r)} &= A - B + (T - 1)C - (T - 1)D \\ &= E - F + (T - 1)G - (T - 1)H.\end{aligned}$$

By (4) and (12),

$$\mathrm{MS}_{(i\times r\times t)} = A - B - C + D = E - F - G + H.$$

Ignoring the mean squares themselves and subtracting the expressions for $\mathrm{MS}_{(i\times r)}$ from corresponding expressions for $\mathrm{MS}_{(i\times r\times t)}$,

$$C - D = G - H.$$

In other words,

$$\overline{\mathrm{cov}\,(X_{rt}, X_{rt'})} - \overline{\mathrm{cov}\,(X_{rt}, X_{r't'})} = \overline{\mathrm{cov}\,(X_{it}, X_{it'})} - \overline{\mathrm{cov}\,(X_{it}, X_{i't'})}.$$

Similarly,

$$B - D = L - M, \text{ and } F - H = K - M.$$

Note in particular the following relationships:

$$B - D = L - M = \frac{1}{R}(y - z) = \frac{1}{R}[\mathrm{MS}_{(i\times t)} - \mathrm{MS}_{(i\times r\times t)}], \tag{25}$$

$$C - D = G - H = \frac{1}{T}(x - z) = \frac{1}{T}[\mathrm{MS}_{(i\times r)} - \mathrm{MS}_{(i\times r\times t)}], \tag{26}$$

and

$$F - H = K - M = \frac{1}{I}(v - z) = \frac{1}{I}[\mathrm{MS}_{(r\times t)} - \mathrm{MS}_{(i\times r\times t)}]. \tag{27}$$

Multirater-Multitrait Matrices

The above formulas (no significance tests implied) pertain to any complete matrix of real numbers, however gathered and regardless of what i, r, and t happen to represent. An especially important application occurs when

i designates ratees, r designates raters, and t designates traits. From the work of Guilford [8], Willingham and Jones [19], and others, the three mean squares involving ratees (r)—MS_r , $MS_{(i\times r)}$, and $MS_{(r\times t)}$—may reflect, respectively, differences among some raters in general level of rating, bias of some raters toward certain individuals, and bias of some raters toward certain traits. Finally, $MS_{(i\times t)}$ reflects differential meaning of the various traits, as Willingham and Jones [19] have shown. "Valid variance" in ratee-rater-trait studies usually consists chiefly of the variance components σ_i^2 and $\sigma_{(i\times t)}^2$; sometimes differences in trait means may be of interest, too.

From the definitions of B, C, and D, one sees that B is the mean of the covariance within traits ($t = t$) across raters ($r \neq r'$). For the tth trait there are $R(R - 1)$ covariances among the R raters, half of these being duplicates of the other half because cov (r, r') = cov (r', r). C is the mean of the covariances within raters ($r = r$) across traits ($t \neq t'$). For the rth rater there are $T(T - 1)$ covariances among the T traits, half of them duplicates of the other half. D, on the other hand, is the mean of the covariances across both raters ($r \neq r'$) and traits ($t \neq t'$). Its magnitude reflects neither interaction of ratees with traits, as does B, nor bias of raters toward ratees, as does C. D constitutes the only internal base for evaluating the magnitudes of B and C. Typically, $B \geq D$ and $C \geq D$, though of course D could exceed B or C. In order to maximize differential meaning of the traits used, B should be as large as possible relative to D, and to minimize the bias of some raters toward certain ratees, $C \leq D$.

Chi ([3], p. 237) sensed part of this latter relationship when he wrote ". . . the correlation between two traits, according to the ratings by one rater, tends to be higher than it should be. On the other hand, since two raters are not likely to take the same attitude or to be under the same prejudice toward an individual rated, the correlation between two traits, according to the ratings by two different raters, would be relatively free from the halo effect. Hence the difference between the former and the latter correlation coefficients may be regarded as the halo effect contained in the ratings by one rater." He performed a factor analysis of such differences and found a general factor of halo, independent of the general factor of the ratings themselves, that accounted for about half as much of the total variance (17 vs. 32 percent).

In a given study, one may find any degree of relative halo effect and any degree of trait independence, for $MS_{(i\times r)}$ is independent of $MS_{(i\times t)}$. These constitute two *separate* criteria for the adequacy of ratings, as Campbell and Fiske [2] and Humphreys [12] point out with respect to multitrait-multimethod matrices. If one reads *rater* in the present paper for *method* in theirs, he has at his command some of the objective summary statistical procedures for which Campbell and Fiske asked.

Formula (25) shows that the difference between B and D (or L and M)

is a simple function of $MS_{(i\times t)}$ and $MS_{(i\times r\times t)}$. B reflects covariation common to all rt, $r't'$ pairings, plus covariation among the rt, $r't$ pairings. $B - D$ is estimated by the final part of (25). For example, (3) may be written as

$$MS_{(i\times t)} = (A - B - C + D) + R(B - D)$$
$$= MS_{(i\times r\times t)} + R(B - D). \tag{28}$$

The "Pigeonhole" Model

How is one to get tests of significance for $B - D$, $C - D$, and $F - H$? From (28), the ratio $MS_{(i\times t)}/MS_{(i\times r\times t)}$ resembles an F ratio, with $(B - D)$ being the effect tested. Is this ratio in fact distributed as F under the null hypothesis? Less stringently, is the right-hand member of (25) an unbiased estimate of the variance component (using that expression in a broad sense) $\sigma^2_{(i\times t)}$? The answer to this latter question would seem to depend upon which analysis-of-variance model yields appropriate expected mean squares for the particular study conducted.

Consider the relatively unrestrictive general linear model set up by Cornfield and Tukey [4] for their "pigeonhole" model (which may also be generalized to an urn-sampling model). By extension of their model for two crossed factors ([4], p. 920), for the rating x_{irts} received by the ith ratee from the rth rater on the tth trait the sth time he is rated by that rater on that trait

$$x_{irts} = \theta + \alpha_i + \gamma_r + \delta_t + \epsilon_{ir} + \eta_{it} + \kappa_{rt} + \lambda_{irt} + \omega_{irts} .$$

Theta represents the general contribution, estimated by $\bar{X}_{\ldots}$ (for the pigeonhole model, $\sigma^2_\theta = 0$). The next seven Greek letters denote the three main contributions and four interactions that are possible. Assumptions are as listed in ([4], pp. 920–921).

Expected mean squares for the finite case of the above linear model are given in ([4], p. 929). (For rather similar E[MS]'s, see [14].) Under what conditions do formulas (25)–(27) estimate variance components without bias?

The right-hand member of (25) estimates the variance component $\sigma^2_{(i\times t)}$ if the raters used in the study were drawn randomly from a large population of raters. The right-hand member of (26) estimates $\sigma^2_{(i\times r)}$ if the traits used in the study were drawn randomly from a large population of traits. The right-hand member of (27) estimates $\sigma^2_{(r\times t)}$ if the ratees used in the study were drawn randomly from a large population of ratees. Otherwise, the respective variance components will tend to be underestimated by formulas (25)–(27), unless $\sigma^2_{(i\times r\times t)} = 0$, as will the analogous F-ratios computed to test significance.

Usually, investigators capture "grab groups" of ratees and raters, who

then constitute the entire population "sampled." Such groups may be composed of volunteers or entire "handcuffed volunteer" classes, but rarely are individuals (ratees or raters) sampled randomly from any defined population. In view of the three conclusions reached above concerning variance components and tests of significance, this appears disturbing. Nearly always we want to generalize beyond the particular ratees and raters used in the study to other ratees and raters "like them." In repeating the study, we would probably use new ratees and raters, but the same traits (though in a given study we might have each ratee rated more than once by each rater on each trait, as allowed for in the above model).

Can we merely *consider* the ratees used in the study as a random sample from a large hypothetical population of ratees "like themselves," and consider the raters similarly? If so, we would have a mixed model (ratees and raters random, traits fixed) for which (25) and (27) would yield unbiased estimates of the variance components $\sigma^2_{(i\times t)}$ and $\sigma^2_{(r\times t)}$.

Cornfield and Tukey ([4], pp. 913–914) tend to encourage this "bootstrap randomization," while Wilk and Kempthorne ([16], pp. 1162–1163; [18], pp. 953–954) discourage it. The latter writers remark: "There are some circumstances under which it may be useful to consider the levels of a random factor actually used as though they were the levels of a fixed factor (with a corresponding redefinition of main effects and interactions), but there appears to be no objective basis for the converse case" ([16], p. 1163).

The matter seems by no means settled yet. By adopting the Cornfield-Tukey point of view we are of course "better off" with the unreplicated ratees-raters-traits study than we would be under the greater restrictions of the Wilk-Kempthorne approach. Replication seems desirable in most instances, however, both within ratee-rater-trait "cells" and across studies with other "grab groups" of ratees and raters. It may be best to assume a fixed-effects model and use $MS_{(i\times r\times s\times t)}$ for testing all effects and interactions in a given study.

Replication within a given study has the added advantage of revealing further biases of raters: $i \times r \times t$, $i \times r \times s$, $r \times s \times t$, and $r \times s$. These can be compensated for statistically in a manner analogous to that of (30), which appears later in this article.

Trait Independence and Rater Bias

B/A should be a close estimate of $\bar{r}_{rt,r't}$, the mean correlation among raters within traits. D/A should be a close estimate of $\bar{r}_{rt,r't'}$, the mean correlation across both raters and traits.

If B significantly exceeds D, then it may be worthwhile to weight the trait scores differentially for predictive purposes. If it does not, then the standard score of the ith individual differs only randomly from trait to trait, and differential weighting is futile. (Here, for the fixed-effects model, we

assume again that $MS_{(i\times r\times t)}$ has as its expected value σ_E^2 , pure error-of-measurement variation [4].)

When statistical significance occurs for $MS_{(i\times t)}$, one may want to find a linear combination of trait factor scores that maximizes the ratio $MS_i/MS_{(i\times r)}$, thereby making differences among the means of individuals as large as possible relative to rater bias toward individuals. This is one way to correct for what Guilford ([8], p. 284) calls relative halo effects. Abelson [1] shows how to employ linear discriminant analysis to maximize variance ratios of this sort. Bias of raters toward ratees is usually so strong that in large studies the $i \times r$ interaction probably shows up as significant, even when $MS_{(i\times r\times t)}$ is used as the error term. Independence of traits and biases of raters toward traits seem less potent.

The better controlled the investigation, the closer D/A will probably approach zero—that is, the poorer the correlation across both raters and traits. (There is, of course, the problem of generally prejudicing information, affecting several raters across traits within ratees.) Careful randomization of the order of presentation of the $I \times T$ ratee-trait combinations, independently for each rater, when experimentally feasible might reduce the extent of interactive rater biases and perhaps increase the independence of traits. (Johnson and Vidulich [13] tried two orders, all traits for one individual vs. all individuals for one trait, but apparently did not randomize anything.)

Consideration of the various possibilities for randomizing the order of ratees, raters, and/or traits used, and of their influences upon expected mean squares, is beyond the scope of this paper; suffice it to say that the analysis of variance mentioned above presupposes *complete* randomization of the order of the $I \times R \times T$ combinations. Kempthorne and collaborators, having contributed greatly to analysis of completely and restrictively randomized designs [16, 17, 18], are now devising analyses (structures) for situations where randomization within the experiment itself can vary from little or none to much or complete, as in the ratee-rater-trait type of investigation. Generally, expected mean squares are considered by them to depend upon what randomization actually takes place within the study (this in addition to the sampling of levels of the factors themselves).

Probably we are well advised to design fuller studies, in which each rater rates each ratee at least twice on each trait. Then there will be a third-order interaction mean square whose mathematical expectation more nearly approaches pure measurement error than does the expected mean square for the second-order interaction.

If this unwillingness to *assume* the variance component for the interaction of ratees, raters, and traits inconsequential seems pedantic, note that we are dealing with two sets of individuals, ratees and raters, organisms probably far more likely to interact with each other and with traits than are

many of the variables manipulated by psychologists. While, for example, strong interaction of style of printing type with size of printing type with color of paper may seem quite unlikely, a priori, we cannot in our present state of ignorance about *intra*-individual characteristics afford to assume that second-order interactions involving individuals are infinitesimal.

Statistical Adjustments for Biases of Raters

Guilford ([8], pp. 280–288) recommends that ratings be adjusted to remove the biases due to raters, reflected in significant MS_r , $MS_{(i\times r)}$, and $MS_{(r\times t)}$. His procedure is equivalent to the following, where X'_{irt} represents the adjusted rating of the ith ratee by the rth rater on the tth trait, and $\bar{X}$'s denote means:

$$X'_{irt} = X_{irt} - (\bar{X}_{.r.} - \bar{X}_{...}) - (\bar{X}_{ir.} - \bar{X}_{i..} - \bar{X}_{.r.} + \bar{X}_{...}) - (\bar{X}_{.rt} - \bar{X}_{.r.} - \bar{X}_{..t} + \bar{X}_{...}). \tag{29}$$

The application of (29) results in adjusted ratings for which MS_r , $MS_{(i\times r)}$, and $MS_{(r\times t)}$ all are zero, but it does not affect $MS_{(i\times r\times t)}$ or the other mean squares. Referring back to (26) and (27), $C - D$ and $F - H$ then become negative: $-MS_{(i\times r\times t)}/T$ and $-MS_{(i\times r\times t)}/I$, respectively. Therefore, Guilford's procedure over-corrects, causing negative bias. The mean covariance across traits within raters becomes less than the mean covariance across traits across raters, representing negative relative halo of magnitude $-MS_{(i\times r\times t)}/T$ when $MS_{(i\times r\times t)}$ is the appropriate error term for $MS_{(i\times r)}$. Similarly, the mean covariance across individuals within traits is made smaller than the mean covariance across individuals across traits.

In order not to over-adjust ratings, one needs a procedure that makes $MS_{(i\times r)}$, $MS_{(r\times t)}$, and MS_r exactly equal to $MS_{(i\times r\times t)}$ without disturbing mean squares other than the three being reduced. This can be done by multiplying each of the two interaction residuals of (29) by the coefficient (1 minus the square root of the ratio of the three-factor interaction mean square to the mean square for the pertinent two-factor interaction): $1 - \sqrt{MS_{(i\times r\times t)}/MS_{(i\times r)}}$ for the first residual and $1 - \sqrt{MS_{(i\times r\times t)}/MS_{(r\times t)}}$ for the second. Also, for the fixed-effects case, multiply $(\bar{X}_{.r.} - \bar{X}_{...})$ by $1 - \sqrt{MS_{(i\times r\times t)}/MS_r}$. Calling these coefficients a, b, and c, respectively, and simplifying, one obtains a formula that makes the nature of the adjusted scores, X''_{irt} , somewhat clearer:

$$X'_{irt} = X_{irt} + a(\bar{X}_{i..} - \bar{X}_{ir.}) + b(\bar{X}_{..t} - \bar{X}_{.rt}) + (a + b - c)(\bar{X}_{.r.} - \bar{X}_{...}). \tag{30}$$

It is easy to show that, by reducing $MS_{(i\times r)}$ to zero, (29) guarantees perfect correlation among raters for total scores of individuals (summed across traits within raters). One estimates the mean correlation among

raters with respect to the sums (or means), over traits, of ratees by ([15], eq. 1)

$$\bar{r}_{X_{ir}.X_{ir'}.} = \frac{\mathrm{MS}_i - \mathrm{MS}_{(i\times r)}}{\mathrm{MS}_i + (R-1)\,\mathrm{MS}_{(i\times r)}}, \tag{31}$$

where $r \neq r'$ and

$$X_{ir.} = \sum^{T} X_{irt} \quad \text{and} \quad X_{ir'.} = \sum^{T} X_{ir't} .$$

For $\mathrm{MS}_{(i\times r)} = 0$, the right side of (31) reduces to $\mathrm{MS}_i/\mathrm{MS}_i$, or unity. Of course the mean r can be unity only when every r between raters is unity. Formula (30) adjusts ratings so as to make $\mathrm{MS}_{(i\times r)}$ equal the originally smaller $\mathrm{MS}_{(i\times r\times t)}$, thereby increasing the average agreement among raters but not rendering it perfect.

Two scores for each ratee are unaffected by the adjustments of formulas (29) and (30):

$$\underline{\sum^{R} X_{irt}} \quad \text{and} \quad \underline{\sum^{R}\sum^{T} X_{irt}} .$$

Therefore, the adjusted trait *sums* (over raters) and adjusted *total* scores (over both raters and traits) cannot be better for any purpose—predictive or otherwise—than the unadjusted ratings were. Furthermore, although the value of $B - D$ in (25) remains constant, both B and D increase equally, while the C of (26) becomes much smaller. In a sense, then, we remove relative halo effect, only to assign it to the general halo effect common to raters without regard to traits.

In fact, the adjustments of (30) typically cause the intercorrelation of the RT rater-trait columns to rise, thereby producing a higher coefficient of equivalence [5] for total scores of individuals across both raters and traits, even though these total scores are not affected at all by the adjustments! This seemingly anomalous result comes about because the adjustment of the $\mathrm{MS}_{(i\times r)}$ downward to the magnitude of $\mathrm{MS}_{(i\times r\times t)}$ increases the numerator of the following formula for the Hoyt-Cronbach [10, 11, 5] coefficient of equivalence, α, without changing the denominator:

$$\alpha_{X_{i..}} = \frac{\mathrm{MS}_i - \mathrm{MS}_{(i\times rt)}}{\mathrm{MS}_i} = \frac{\mathrm{MS}_i - \dfrac{\mathrm{SS}_{(i\times r)} + \mathrm{SS}_{(i\times t)} + \mathrm{SS}_{(i\times r\times t)}}{(I-1)(RT-1)}}{\mathrm{MS}_i}, \tag{32}$$

where

$$X_{i..} = \sum^{R}\sum^{T} X_{irt}$$

and where the SS's are sums of squared deviations (i.e., mean squares multiplied by their respective degrees of freedom). Thus the statistical adjust-

ment for relative halo affect cannot affect test-retest or comparable-forms reliability, though when positive halo exists, it does increase the *estimated* internal consistency. (To understand this formula better, see (34) in the Appendix.)

The above paradox arises because one is dealing with a test of the sort that Cronbach [5] calls "lumpy," and also because one treats the new $MS_{(iXrt)}$ as if it still had $(I - 1)(RT - 1)$ degrees of freedom, when in reality it now has only $(I - 1)(RT - R)$ d.f., because by setting $MS_{(iXr)}$ at a fixed value—that of $MS_{(iXrXt)}$—one loses $(I - 1)(R - 1)$ d.f. The reduction in d.f. may or may not compensate for the reduction in the magnitude of $MS_{(iXr)}$, so the alpha of (32) might change in either direction. Usually its magnitude will increase.

Though one may have uses for ratings adjusted by (30), such statistical manipulations should by no means substitute for careful designing of the rating study to minimize bias and maximize independence of traits experimentally. Typically, experimental control is superior to statistical control. Where the latter is needed also, Abelson's procedure [1] for finding factors in the traits that maximize $MS_i/MS_{(iXr)}$ may, when there is significant interaction of ratees with traits, be preferable to (30).

If one had better estimates of error than $MS_{(iXrXt)}$, he should use them, instead, for obtaining the a, b, and c that appear in (30). When significant second-order $(i \times r \times t)$ interaction occurs, (30) may adjust too little, this depending upon the appropriate analysis-of-variance model. If each rater rated each ratee $S > 1$ times on each trait, one might employ $MS_{(iXrXsXt)}$, rather than $MS_{(iXrXt)}$, for securing a, b, and c, again depending upon the relevant model.

A Numerical Example

Consider Guilford's individuals-raters-traits data ([8], pp. 282–288) from the above point of view. There were 105 ratings, with $I = 7$, $R = 3$, and $T = 5$. Table 1 contains the various mean squares and tests of significance. All main effects and interactions except $MS_{(rXt)}$ are significantly larger than $MS_{(iXrXt)}$ beyond the .05 level.

Applying formulas (5) through (8), $A = 3.351$; $B = 0.763$, $B/A = .23$; $C = 1.851$, $C/A = .55$; and $D = 0.443$, $D/A = .13$. The .23 is identical with the comparable item in Guilford's Table 11.6, and the .55 is almost identical with the mean of the .70, .25, and .74 in the last column of his Table 11.7.

From (2), $MS_{(iXr)} = (A - B - C + D) + T(C - D) = 8.22$, highly significant when compared with $MS_{(iXrXt)} = A - B - C + D = 1.18$ because of the large covariance among traits within raters (C) compared with the small covariance across both raters and traits (D). The mean of the 30 intra-rater coefficients of correlations among traits, estimated by C/A, was

TABLE 1

Analysis of Variance of Ratings of Seven Ratees by Three Raters on Five Traits, after Guilford ([8], p. 283)*

Source of Variation	*d.f.*	Mean Square	MS/1.18	P
Among ratees (i)	6	15.82	13.41	$<.001$
Among raters (r)	2	4.52	3.83	$<.05$
Among traits (t)	4	11.63	9.86	$<.001$
$i \times r$	12	8.22	6.96	$<.001$
$i \times t$	24	2.14	1.81	$<.05$
$r \times t$	8	1.62	1.37	$>.05$
$i \times r \times t$	48	1.18	—	—
Total	104	3.56	—	—

*But, using a different procedure for testing significance, Guilford failed to find r or $(i \times t)$ significant.

.55, contrasted with a D/A of only .13. Clearly, strong relative halo effect occurred in this study.

Similarly but less markedly, $MS_{(i\times t)} = (A - B - C + D) + R(B - D) = 2.14$, significant at the .05 level. The average of the 15 intercorrelations among raters within traits was estimated by B/A to be .23, contrasted with the base-line $\bar{r}$ of .13. Therefore, the traits are to some extent different, though probably not as much as the investigators desired. Finally, $MS_{(r\times t)}$ is not significant; from formulas (13), (14), and (16) one can estimate, via F/E, that the mean correlation across ratees within traits is $-.03$, contrasted with $-.06$ for the $\bar{r}$ across both ratees and traits, estimated by H/E.

Reducing $MS_{(i\times r)}$ to the magnitude of $MS_{(i\times r\times t)}$ via the adjustment in (30) changes B/A from .23 to .51, C/A from .55 to .38, and D/A from .13 to .38. The apparent gain in trait independence is spurious, of course, because both $MS_{(i\times t)}$ and $MS_{(i\times r\times t)}$ are unaltered; the $\sum^R X_{irt}$'s are unaffected by the adjustments. Relative halo effect did disappear, being absorbed into the base-line correlation across both raters and traits, reflected by the considerable rise in D/A.

The average of the three r's among raters, estimated by means of (31), changes from .24 for the original rating sums, $\sum^T X_{irt}$, to .81 among such sums of ratings adjusted by (30). The coefficient of equivalence rises from .84 for unadjusted ratings to .89 or .91 for adjusted ones, depending upon how many degrees of freedom, $(I - 1)(RT - R)$ or $(I - 1)(RT - 1)$, are used in (32).

An Extension

For many analysis-of-variance situations one needs a mean square whose mathematical expectation is just σ^2, or very nearly so, in order to devise proper error terms and to estimate components of variance. Having each rater rate each ratee-trait combination more than once under randomized conditions that minimize memory carryover will help meet this need. The multiple ratings of each ratee on each trait can be considered an ordered fourth (fixed?) effect, say sequence, with $s = 1, 2, \cdots, S$; $S > 1$. Now the notation for the rating received by the ith ratee from the rth rater the sth time on the tth trait is X_{irst}. If the $\mathrm{MS}_{(i\times r\times s\times t)}$ has a relatively large number of degrees of freedom, it might be employed as the MS with $E[\mathrm{MS}] = \sigma^2$, under the reasonable assumption that $\sigma^2_{(i\times r\times s\times t)}$, the component of variance attributable to the third-order interaction, is negligible.

A complete analysis of such ratings, both by analysis-of-variance and correlational methods, may be worthwhile, especially for such comparisons as $\bar{r}_{rst,rs't}$ with $\bar{r}_{rst,r'st}$ to check upon intra-rater versus inter-rater reliability. Components of variance should also be informative. If $S > 2$, one might employ orthogonal polynomials to test for nonlinear trends in the rating sequence [7].

For the four-factor design there are seven mean covariances, as contrasted with three for the three-factor design; these are

$$\overline{\mathrm{cov}\,(X_{rst}, X_{r'st})}, \quad \overline{\mathrm{cov}\,(X_{rst}, X_{rs't})}, \quad \cdots, \quad \overline{\mathrm{cov}\,(X_{rst}, X_{r's't'})}.$$

Because eight mean squares involve ratees, the seven mean covariances and $\bar{s}^2_{rst}$ can be computed.

Concluding Remark

It seems quite likely that the formulas given here are applicable far beyond the ratee-rater-trait situation. Abelson's heuristic table [1] classifying agents, objects, and modes for six types of studies lists the following possibilities from sociometry, clinical ratings, the semantic differential, laboratory experiments, psychological testing, and psychophysical or preference ratings: judges-judgees-items, raters-concepts-scales, conditions-subjects-responses, subjects-conditions-responses (trials?), occasions-subjects-tests, and judges-stimuli-(hypothetical) scale components.

Perhaps approaching a three-way classification of real numbers in the ways suggested in this paper furthers Abelson's goal of offering "a promising combination of experimental and correlational approaches" and partially resolves the dilemma to which Cronbach [6] pointed.

Appendix: Outline of Proof

Gulliksen ([9], p. 54) and Stanley ([15], pp. 90–91) have shown that the $\mathrm{MS}_{(i\times j)}$ of a two-way classification is equivalent to $\bar{s}^2_j - \overline{\mathrm{cov}\,(X_j, X_{j'})}$

where $j \neq j'$. Applying this relationship to the matrix of individuals-by-trait means (over raters), one can by the following procedure secure formula (3):

$$\mathrm{MS}_{(i\times t)} = \sum^{I}\sum^{R}\sum^{T}(\bar{X}_{i\cdot t} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot\cdot t} + \bar{X}_{\cdot\cdot\cdot})^2/(I-1)(T-1)$$

$$= R\sum^{I}\sum^{T}\left[\bar{X}_{i\cdot t} - \frac{\sum^{I}\bar{X}_{i\cdot t}}{I} - \frac{\sum^{T}\bar{X}_{i\cdot t}}{T} + \frac{\sum^{I}\sum^{T}\bar{X}_{i\cdot t}}{IT}\right]^2/(I-1)(T-1)$$

$$= R\left[\frac{\sum^{T} s^2_{(\sum^{R} X_{irt}/R)}}{T} - \frac{\sum^{T}\sum^{T-1}\mathrm{cov}\,(\sum^{R} X_{irt}/R\,,\ \sum^{R} X_{irt'}/R)}{T(T-1)}\right]$$

$$= \left[\sum^{T} s^2_{(X_{1t}+X_{2t}+\cdots+X_{Rt})} - \frac{\sum^{T}\sum^{T-1}\mathrm{cov}\,(X_{1t}+\cdots+X_{Rt}\,,\ X_{1t'}+\cdots+X_{Rt'})}{T-1}\right]/RT$$

$$= \left\{\sum^{T}\left[\sum^{R} s^2_{rt} + \sum^{R}\sum^{R-1}\mathrm{cov}\,(X_{rt}\,,\ X_{rt'})\right] - \sum^{T}\sum^{T-1}\left[\sum^{R}\mathrm{cov}\,(X_{rt}\,,\ X_{rt'}) + \sum^{R}\sum^{R-1}\mathrm{cov}\,(X_{rt}\,,\ X_{rt'})\right]/(T-1)\right\}/RT$$

$$= \overline{s^2_{rt}} + (R-1)\,\overline{\mathrm{cov}\,(X_{rt}\,,\ X_{rt'})} - \overline{\mathrm{cov}\,(X_{rt}\,,\ X_{rt'})} - (R-1)\,\overline{\mathrm{cov}\,(X_{rt}\,,\ X_{rt'})} = A + (R-1)B - C - (R-1)D.$$

Formulas for the other first-order interactions can be obtained in the same way as (3), above.

To secure (4), for $\mathrm{MS}_{(i\times r\times t)}$,

$$\mathrm{MS}_{(i\times r\times t)} = A - [(R-1)B + (T-1)C + (R-1)(T-1)D]/(RT-1) \tag{33}$$

and then

$$(\text{Sum of Squares})_{(i\times r\times t)} = \mathrm{SS}_{(i\times r)} + \mathrm{SS}_{(i\times t)} + \mathrm{SS}_{(i\times r\times t)}\,. \tag{34}$$

Formulas (1), (9), and (17) are readily secured in a straightforward manner from the definitional formulas for MS_i, MS_r, and MS_t. Finally, note that, for example,

$$RT(I-1)A = \mathrm{SS}_i + \mathrm{SS}_{(i\times r)} + \mathrm{SS}_{(i\times t)} + \mathrm{SS}_{(i\times r\times t)}\,. \tag{35}$$

This relationship, known from fundamental considerations of the analysis of variance *before* solving for A via formulas (1)–(4), constitutes an independent check of (5) and, therefore, indirectly of (6)–(8).

REFERENCES

[1] Abelson, R. P. A discriminant approach to factoring three-way data tables. *Amer. Psychologist*, 1958, **13**, 375. (Abstract) (More extensive reports privately circulated.)
[2] Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81–105.
[3] Chi, P.-L. Statistical analysis of personality rating. *J. exp. Educ.*, 1937, **5**, 229–245.
[4] Cornfield, J. and Tukey, J. W. Average values of mean squares in factorials. *Ann. math. Statist.*, 1956, **27**, 907–949.
[5] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.
[6] Cronbach, L. J. The two disciplines of scientific psychology. *Amer. Psychologist*, 1957, **12**, 671–684.
[7] Grant, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, **53**, 141–154.
[8] Guilford, J. P. *Psychometric methods* (2nd ed.) New York: McGraw-Hill, 1954.
[9] Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.
[10] Hoyt, C. J. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, **6**, 153–160.
[11] Hoyt, C. J. and Stunkard, C. L. Estimation of test reliability for unrestricted item scoring methods. *Educ. psychol. Measmt*, 1951, **12**, 756–758.
[12] Humphreys, L. G. Note on the multitrait-multimethod matrix. *Psychol. Bull.*, 1960, **57**, 86–88.
[13] Johnson, D. M. and Vidulich, R. N. Experimental manipulation of the halo effect. *J. appl. Psychol.*, 1956, **40**, 130–134.
[14] Stanley, J. C. Fixed, random, and mixed models in the analysis of variance as special cases of a finite model. *Psychol. Rep.*, 1956, **2**, 369.
[15] Stanley, J. C. K-R 20 as the stepped-up mean item intercorrelation. *14th Yrbk Natl Coun. Meas. used in Educ.*, 1957. Pp. 78–92.
[16] Wilk, M. B. and Kempthorne, O. Fixed, mixed, and random models. *J. Amer. statist. Ass.*, 1955, **50**, 1144–1167.
[17] Wilk, M. B. and Kempthorne, O. Derived linear models and their use in the analysis of randomized experiments. *WADC Tech. Rep.* 55-244, Vol. II, Mar. 1956.
[18] Wilk, M. B. and Kempthorne, O. Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. math. Statist.*, 1956, **27**, 950–985.
[19] Willingham, W. W. and Jones, M. B. On the identification of halo through analysis of variance. *Educ. psychol. Measmt*, 1958, **18**, 403–407.

Manuscript received 11/20/59
Revised manuscript received 10/21/60