THE PREDICTION OF COLLEGE GRADES FROM COLLEGE BOARD SCORES AND HIGH SCHOOL GRADES

by

Richard F. Potthoff

University of North Carolina

Institute of Statistics Mimeo Series No. 419

December 1964

The adjustment of high school grades for differences among high schools in grading standards and practices is a problem which must be dealt with in any prediction system which utilizes high school grades in predicting academic success in college. Similarly, the need for adjustment of college grades for differences among colleges will also enter into the picture in the operation of a central prediction system. This report examines and evaluates different possible approaches for the prediction of college grades using College Board scores and high school grades as the predictors, considers what assumptions are made under the different approaches, and obtains the maximum-likelihood estimators of the parameters which are used to adjust the high school grades and college grades under a large number of different possible models. The variances of the predicted college grades, as well as confidence intervals associated with these predicted grades, are considered.

This research was supported by Educational Testing Service, and was also supported in part by the Mathematics Division of the Air Force Office of Scientific Research.

DEPARTMENT OF STATISTICS

UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

TABLE OF CONTENTS

THE PREDICTION OF COLLEGE GRADES FROM COLLEGE BOARD SCORES AND HIGH SCHOOL GRADES*

by

Richard F. Potthoff

University of North Carolina

1. Introduction and summary. Certain statistical problems will arise in any program to set up a central prediction system under which the college grades (C) of a large number of students are to be predicted using both their College Board test scores (T) and their high school grades (H) as predictors. The purpose of this report is to explore different possible statistical methods and models for effecting such predictions, and to make appropriate evaluations of the different alternatives.

A major reason for the existence of statistical complications in this area is the fact that grades in one high school cannot be assumed to be equivalent to grades in any other high school and grades in one college likewise cannot be assumed to be equivalent to grades in another college. Thus something must be done which will result in equating the grades from the different high schools and equating the grades from the different colleges. How to effect this equating is a basic statistical problem which must be faced in connection with the establishment of a central prediction system.

One idea which has been suggested is to equate the college grades by using the College Board test scores, and then, once all the college grades are on a common basis so that they are comparable, the high school grades can be

*This research was supported by Educational Testing Service, and was also supported in part by the Mathematics Division of the Air Force Office of Scientific Research.

equated through the use of standard methods. A second suggestion more or less reverses this process: the grades of different high schools are equated on the basis of the College Board test scores, and then these equated high school grades are used (along with the test scores) for equating the college grades. Still a third possibility is to use a technique which does the equating for the high school grades and for the college grades simultaneously.

All three of these possible approaches will be considered in some detail (see Sections 5, 7, and 9), and certain basic models will be formulated and studied. For all of the formulations, we will show how to obtain the maximum-likelihood estimates of the equating parameters.

All three approaches will be evaluated (see Sections 6, 8, and 10). The third one is tentatively recommended over the other two, since it would seem to be less vulnerable to sneaky systematic biases by virtue of the fact that the assumptions upon which it is based appear to be more realistic and less restrictive. As an alternative choice, a special form of the second approach is recommended, for there might exist certain conditions under which this choice could result in a bit more efficient estimation and prediction than the third approach, and also the calculations are noticeably simpler. The third approach requires much lengthier computations than the other two; although the recommendation in its favor is contingent (for one thing) upon the feasibility of performing these computations, the equation systems to be solved have been carefully examined and it appears that their solution is entirely practicable provided that a little time can be obtained on a sufficiently large computer.

Once the maximum-likelihood estimates of the equating parameters have been found, the predicted college grades can be obtained. It might be desirable to obtain for each student not only his predicted college grade, but also

a confidence interval for this predicted grade. These matters are covered after the discussion of the third approach (see Section 11). The confidence intervals on the predicted grades will tend to be narrower for the students from the larger high schools, since the estimation of the equating parameters for a high school will tend to be more accurate the larger the high school.

This report consists of two parts. The second part is a Mathematical Appendix (consisting of a series of notes) to which we have relegated some of the more technical details and mathematical derivations. Here in the first part, which is less technical, we present and discuss the principal findings of the report, and at the same time we exhibit the important formulas. All of the reference numbers ([1], [2], etc.) here in the first part will refer to the notes in the Mathematical Appendix.

2. Notation, assumptions, and understandings. We will be concerned essentially with three types of variables: college grades (C), College Board test scores (T), and high school grades (H). The object of the central prediction system will be to predict the value of the variable C for each of a large number of students for whom the values of T and H are given. This predicting is to be done by examining data on all three variables from an earlier group of students for whom data on C (as well as on T and H) is already available, and then finding an optimal way of predicting C on the basis of T and H. For example, students who graduated from high school in 1964 will have received some college grades by early 1965; the data on C, T, and H from these students could be employed to develop a way of predicting C given T and H, to be used on students who will graduate in 1965 and for whom only T and H data will be available.

Most of our development will deal with the earlier group (i.e., the

group for which C data as well as T and H data is available), since it is this group which is utilized to estimate the parameters of the prediction system which is to be applied to the later group. With respect to the earlier group in particular, we adopt the following notation. Let $N..$ be the number of students in the group. Let them be distributed among n different colleges, and let m be the number of different high schools from which these $N..$ students came. Let N_{ij} denote the number of students in the j-th college who came from the i-th high school. We also define $N_{.j}$ to be the number of students in the j-th college, and $N_{i.}$ to be the number of students who came from the i-th high school. Thus

$$N_{i.} = \sum_{j=1}^{n} N_{ij}\ ,\ N_{.j} = \sum_{i=1}^{m} N_{ij},\ \text{and}\ N.. = \sum_{i=1}^{m} N_{i.} = \sum_{j=1}^{n} N_{.j} = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\ .$$

Out notation will identify each student by means of a triple of indices (i,j,k), where i refers to the high school from which he came (i = 1, 2, ..., m), j refers to the college which he is attending (j = 1,2,...,n), and the index k ($k=1,2,\ldots,N_{ij}$) is used to distinguish the N_{ij} different students who are in the j-th college and came from the i-th high school. Then C_{ijk}, T_{ijk}, and H_{ijk} will represent respectively the college grade average, College Board test score, and high school grade average of the student with identification (i, j, k); C_{ijk} is in terms of the grading scale of college j, and H_{ijk} is in terms of the grading scale of high school i. We also define

$$(2.1)\quad C_{ij} = \sum_{k=1}^{N_{ij}} C_{ijk},\ C_{i.} = \sum_{j=1}^{n} C_{ij},\ C_{.j} = \sum_{i=1}^{m} C_{ij},\ C.. = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}\ ,$$

$$\bar{C}_{ij} = C_{ij}/N_{ij}\ ,\ \bar{C}_{i.} = C_{i.}/N_{i.}\ ,\ \bar{C}_{.j} = C_{.j}/N_{.j},\ \bar{C}.. = C../N..\quad .$$

By substituting the letter T for the letter C wherever the latter appears in (2.1), we define eight more expressions; likewise, by substituting H for C, we define another eight expressions.

Perhaps our terminology should be explained more precisely. The term "high school" embraces what might otherwise be called "secondary schools", "preparatory schools", or simply "schools". The term "college" of course embraces "universities". The term "college grade" (C_{ijk}) refers to a college grade average over some specified period of time (or possibly to some closely related measurement); e.g., C_{ijk} might be the average of all the individual's grades for the entire freshman year. By "high school grade" (H_{ijk}) we mean the average of the student's high school grades over a designated period of time (which might be anywhere from one to four years), or possibly some related measurement such as one based on his rank in class. By "College Board test score" (T_{ijk}) we mean a single score (which might be a sum or weighted combination of other scores) received by the student in a common testing program which was administered to all $N_{..}$ students.

We will not consider the possibility of using more than one H-variable as a predictor (e.g., one might attempt to use high school grade averages in several different subject areas as predictors, instead of using the single predictor consisting of the over-all grade average). We likewise will not consider the possibility of trying to predict more than one C-variable (e.g., one could try to set up a system with multiple criterion variables consisting of college grades in several different subject areas, rather than restricting oneself to the single over-all college grade average). Such refinements as these would complicate the statistical analysis of the system considerably, and so at this stage in development it appears best to concentrate on the more basic models. We will, however, consider at various points in this report

the use of more than one T-variable (i.e., more than one type of test score arising from a common testing program administered to all N.. students) as a predictor, since such a generalization does not cause as much complication.

For some purposes, such as judging the extent and cost of the computation that will be required with different prediction schemes, it will be helpful to know the approximate values of m and n. It is anticipated that m, the number of high schools encompassed in the system, may be as high as 4000 or 5000, and that n, the number of colleges in the system, will be roughly 400 or 500.

For carrying out the equating for the college grades and for the high school grades, we will assume that a linear transformation of the grades of each college or each high school will be a sufficiently general type of transformation. Thus we associate with college j a pair of constants α_j and β_j, so determined that the variable

$$c_{ijk} = \alpha_j + \beta_j C_{ijk} \tag{2.2}$$

represents an "equated" college grade; in other words, two c_{ijk}'s (2.2) from any two different colleges are always comparable, whereas C_{ijk}'s from different colleges are of course not comparable. Similarly, we associate with each high school a pair of equating parameters a_i and b_i such that the variable

$$h_{ijk} = a_i + b_i H_{ijk} \tag{2.3}$$

represents an equated high school grade, and thereby compensates for differences in grading standards among the high schools. In the models which we will treat, a regression parameter will ordinarily be considered to be absorbed in the a_i's and b_i's.

Actually, the α_j's, β_j's, a_i's, and b_i's (as well as a couple of other parameters) are all unknown. It is the estimation of these parameters which constitutes our principal statistical problem.

With the c_{ijk}'s, the C_{ijk}'s, the h_{ijk}'s, the H_{ijk}'s, and the T_{ijk}'s, we will adopt the convention that these variables are all scored in such a way that a better performance is reflected in a higher (rather than a lower) value of the variable. This implies, incidentally, that all β_j's [see (2.2)] and all b_i's [see (2.3)] are > 0.

If we wish, we can use only one rather than two parameters for equating, and eliminate the term β_j in (2.2) and/or the term b_i in (2.3). The resulting set-ups would be less complicated, but at the same time less general and presumably less accurate. We shall consider such set-ups at various points in this report, however.

Although we will not introduce all of our notation at this stage, we define the following expressions before closing this section:

$$(2.4) \qquad S_{HHi.} = \sum_j \sum_k H^2_{ijk} - N_{i.} \bar{H}^2_{i.}$$

$$S_{THi.} = \sum_j \sum_k T_{ijk} H_{ijk} - N_{i.} \bar{T}_{i.} \bar{H}_{i.}$$

$$S_{TTi.} = \sum_j \sum_k T^2_{ijk} - N_{i.} \bar{T}^2_{i.}$$

$$S_{TT} = \sum_i \sum_j \sum_k T^2_{ijk} - N_{..} \bar{T}^2_{..}$$

$$S_{CC.j} = \sum_i \sum_k C^2_{ijk} - N_{.j} \bar{C}^2_{.j}$$

$$S_{CT.j} = \sum_i \sum_k C_{ijk} T_{ijk} - N_{.j} \bar{C}_{.j} \bar{T}_{.j}$$

$$S_{TT.j} = \sum_i \sum_k T^2_{ijk} - N_{.j} \bar{T}^2_{.j}$$

$$r_j = S_{CT.j} / (S_{CC.j} S_{TT.j})^{\frac{1}{2}}$$

$$S_{CTij} = \sum_k C_{ijk} T_{ijk} - N_{ij} \bar{C}_{ij} \bar{T}_{ij}$$

$$S_{CHij} = \sum_k C_{ijk} H_{ijk} - N_{ij} \bar{C}_{ij} \bar{H}_{ij}$$

$$d_{ij} = H_{ij} - N_{ij} \bar{H}_{i.}$$

$$e_{ij} = T_{ij} - N_{ij} \bar{T}_{i.}$$

3. Previous related studies. We consider briefly some previous work which is related to the material of this report. Tucker [9], after mentioning some earlier work in the areas of equating of grades and prediction of college grades, examines three different models for a central prediction system. The first two are canonical correlation models, which Tucker himself does not favor for prediction purposes. The third one, which is called the "predictive model", is extremely general (as are the other two), and, in fact, is more general than our formulations in a number of respects, including the allowance for provisions to take care of more than one C-variable and more than one H-variable; with this greater generality, however, are associated more serious computational complications. At the same time, Tucker's third model seems to be less general than our formulations in one respect: it apparently makes no provision for anything similar to our β_j's (see (2.2)), and so a problem develops as to how to weight certain error terms (see [9], p. 55).

As Tucker [9, p.2] points out, one method which has been used to predict college grades is for a given college to use data on its own current students to set up a regression system for predicting the grades of its prospective students. This method is not only computationally simple but also rests on a minimum of assumptions. However, its fault lies in the fact that it uses only a relatively small amount of data, which causes the estimates of

the equating parameters, and consequently also the predictions, to be comparatively inefficient and inaccurate; in fact, the grades from those high schools which send only a few students to the college can hardly be used at all in any prediction, since the estimates of the equating parameters for such high schools would be so unstable. With a central prediction system, on the other hand, the entire mass of data encompassing all n colleges, all m high schools, and all $N_{..}$ students is utilized simultaneously to estimate the various equating parameters, and this should result in much better estimation and improved predictions.

Gulliksen, in a section entitled "Equating two forms of a test given to different groups" [5, p. 299ff.], presents a theoretical development (which he attributes to Tucker) that seems to be applicable to the problem of equating college grades (C) of different colleges on the basis of a common test (T) administered to all students. The development treats explicitly the special case which would correspond to only $n = 2$ colleges, but generalization to $n > 2$ is immediate. In any event, though, the development deals essentially with the population parameters themselves rather than with the estimation of these parameters, whereas in this report all of our methodology will be based on maximum-likelihood estimation of unknown parameters. The Gulliksen-Tucker development assumes that there has been selection on the basis of the equating variable (T) and only on the basis of T (i.e., not on C or on some third variable such as H); the case where there is selection on the basis of the variable to be equated rather than on the equating variable (such as might be assumed to exist if T is the equating variable and H the variable to be equated, e.g.) is not considered by Gulliksen, but simlar tools might be applied to this case to obtain relations between the population parameters.

The effect of selection receives considerable attention in Gulliksen's

book [5] in various connections; it turns out that the sometimes tricky influence of selection will also make itself felt in various phases of this report. One basic principle which we will be relying upon is the fact that, if X and Y are two variables such that selection is made on X but not on Y, then the conditional distribution of Y given X is unaltered by the selection on X (whereas the joint distribution of X and Y is not generally unaltered, and neither is the conditional distribution of X given Y).

4. The use of different prediction equations for different groups. In his study, Tucker [9] goes to some length to provide for the possibility of using different regression equations, or predictive composites, for the grade predictions for different groups of colleges. In particular, for example, he explores a system in which grades at liberal arts colleges are predicted via one set of regression weights, and grades at engineering and technical colleges are predicted via a second set of regression weights. The argument is that the grades at the two different types of colleges essentially constitute two different types of criterion variables, so that it is more realistic to have two separate sets of regression weights for predicting them.

Such refinements as these have the advantage of generalizing the basic model, but at the same time they create certain complications. The computations seem to be comparatively cumbersome in relation to the sort of computational requirements that are proposed here in this report. Also, Tucker [9, see especially pp. 54-55] takes note of several unsolved mathematical problems which arise with his "predictive model". The latter include questions as to the uniqueness of the solution for the estimates of the parameters, as well as, perhaps, the convergence of the iterative process leading to this solution; the choice of a certain integer which he calls n_p (which is the number

of predictive composites, and is ≥ 1 but $\leq n$, the number of colleges); and the weighting of the squared errors. It appears that, at the present stage of development, it might be best, so far as practical application is concerned, not to move right away into a relatively sophisticated model which allows for splitting the colleges into groups with different basic prediction equations for the different groups. Nevertheless, though, a model like Tucker's provides stimulating food-for-thought for future stages of development.

In this report, we will assume that the criterion variable (college grade average) is basically the same variable for all n colleges (but with linear transformations being required to equate grades of different colleges), so that for all n colleges the same regression weights can be applied to the predictor variables. This assumption would seem to be considerably more reasonable for freshman grades than for post-freshman grades, since curriculums would tend to become less homogeneous the more advanced the students are in college. In case it is felt, however, that the regression weights for the engineering (technical) colleges really should be different from those for the liberal arts colleges, then the engineering colleges (which would probably not account for too large a fraction of the total) could be taken as excluded from the group of n colleges. The estimates of the a_i's and b_i's based on the students at the n liberal arts colleges might then be utilized somehow in setting up a prediction system for the (presumably) small number of engineering colleges.

Finally, we mention a type of grouping which is different from grouping by colleges. The question might arise as to whether we should have one set of regression weights for predicting girls' grades and a second set for predicting boys' grades. It is to be hoped that it will be satisfactory to assume the same set for both, or that, at worst, the appending to the model of a

single extra predictor (essentially a second T variable, but able to assume only two values) will suffice to take care of any sex differences.

5. The first approach: equating the college grades by means of the T variable, followed by equating of the high school grades. We are now ready to turn to a detailed consideration of the three approaches to equating which were mentioned in Section 1. In the first approach, the first step is to equate the college grades using the T-scores as the equating variable. We will assume a model in which the conditional distribution of c_{ijk} (2.2) given T_{ijk} is normal with variance equal to 1 and with mean equal to a linear function of T_{ijk}. Such a model, which we will designate by $C \mid T$, seems to be the most realistic (i.e., more realistic than a model which specifies a bivariate normal distribution for C and T, or one which specifies that the conditional distribution of T given C is normal), since there certainly is selection on the basis of T whereas there is no selection based on C. If the joint distribution of C and T before selection on T was bivariate normal, then the conditional distribution of C given T would be unaltered by selection on T, i.e., it would be normal with the same mean and variance both before and after the selection.

Under our model which we just specified, the conditional distribution of c_{ijk} given T_{ijk} may be written in the form

$$(5.1) \qquad (2\pi)^{-\frac{1}{2}} \; e^{-\frac{1}{2}(c_{ijk} - \mu - \nu T_{ijk})^2} \quad ,$$

where μ and ν are unknown parameters (and ν will be > 0 because of positive correlation between college grades and test scores). It is the C_{ijk}'s rather than the c_{ijk}'s which are the observed quantities, however. If we apply the transformation

(5.2) $$c_{ijk} = \alpha'_j + \beta_j C_{ijk}$$

to (5.1), then, after first getting

(5.3) $$\left|\frac{d\, c_{ijk}}{d\, C_{ijk}}\right| = |\beta_j|$$

from (5.2), we use (5.1 - 5.3) to find that the conditional distribution of C_{ijk} given T_{ijk} is

(5.4) $$(2\pi)^{-\frac{1}{2}} |\beta_j| \; e^{-\frac{1}{2}(\alpha_j + \beta_j C_{ijk} - \nu T_{ijk})^2} \quad ,$$

where

(5.5) $$\alpha_j = \alpha'_j - \mu \quad .$$

For convenience, we denoted the additive term by α'_j in (5.2) [rather than by α_j as in (2.2)], so that we could save the α_j notation for (5.5) which absorbs both α'_j and μ. Note also that the assumption about the variance being equal to 1 is not really an arbitrary assumption, but rather it still allows for complete generality since, in effect, we can think of the standard deviation parameter as being absorbed in the α_j's, the β_j's, and ν in (5.4); or, to look at it another way, we may observe that the formulation (2.2) [or (5.2)] is not actually unique (since it will still be valid if the α_j's and β_j's are multiplied by any positive constant), and the assumption that the variance is 1 can simply be thought of as a means of making the formulation unique.

It is (5.4) which we use in getting the maximum-likelihood estimates of the α_j's and β_j's (and also, incidentally, of ν). Thanks to the complicating presence of the β_j's, the obtaining of these maximum-likelihood estimates

in this case is not merely a standard problem in least-squares estimation involving simply the solution of a linear equation system. What we have to do is to take the logarithm of the product over i,j,k of the expressions (5.4); differentiate it with respect to the α_j's, the β_j's, and ν; set the resulting derivatives equal to 0; and solve this equation system for the α_j's, the β_j's, and ν. The formulas for the estimates (the details of deriving which are given [1] in the Mathematical Appendix) are as follows. First we solve the equation

$$(5.6) \qquad \sum_j S_{TT.j}\left[2-r_j^2-r_j^2\left(1+\frac{4N_{.j}}{S_{TT.j}r_j^2\nu^2}\right)^{\frac{1}{2}}\right] = 0$$

for ν^2, by using (e.g.) the Newton-Raphson method[2]. We take the positive square root of this solution ν^2 to get $\hat{\nu}$, the estimate of ν. Then we calculate

$$(5.7) \qquad \hat{\beta}_j = \frac{\hat{\nu}S_{CT.j}}{2S_{CC.j}}\left[1+\left(1+\frac{4N_{.j}}{S_{TT.j}r_j^2\hat{\nu}^2}\right)^{\frac{1}{2}}\right]$$

and

$$(5.8) \qquad \hat{\alpha}_j = \hat{\nu}\bar{T}_{.j} - \hat{\beta}_j\bar{C}_{.j} \quad ,$$

the estimates of the β_j's and α_j's. Note that $\hat{\beta}_j$(5.7) will be > 0 so long as

$$(5.9) \qquad S_{CT.j} > 0 \quad .$$

Now (5.9) merely requires C and T to have a positive sample correlation for each college, a condition which must certainly be satisfied if the grades of the college are to have any meaningful relation to T at all; in the unlikely event that $S_{CT.j}$ is < 0 for some college, such a college should probably be thrown out of the system anyway.

We might wish to consider the problem of how to equate college grades when there is more than one T-variable upon which to base the equating. The maximum-likelihood estimates can still be obtained for this case, but the calculations may be noticeably more complicated than those which are associated with the relatively simple formulas (5.6 - 5.8, A2.2). Some details for this case are given in the Appendix[3].

We observe that, if all β_j's are eliminated (i.e., set equal to 1) in (5.2-5.4), then (5.4) becomes

$$(5.10) \qquad (2\pi\sigma^2)^{-\frac{1}{2}} \; e^{-\frac{1}{2}(C_{ijk} + \alpha_j - \nu T_{ijk})^2/\sigma^2}$$

after including a parameter σ^2 for the variance. Thus (5.10) represents a simplified model with only one equating parameter (α_j) instead of two (α_j, β_j) for each college. The estimation of α_j under the model (5.10) is nothing but a standard problem in least squares analysis: the estimator is

$$(5.11) \qquad \hat{\alpha}_j = \hat{\nu}\,\overline{T}_{\cdot j} - \overline{C}_{\cdot j} \quad ,$$

where

$$(5.12) \qquad \hat{\nu} = \sum_j S_{CT.j} / \sum_j S_{TT.j} \quad .$$

This completes the discussion of the first step (equating of college grades via the T-scores) of our first approach to equating. We now consider the equating of the high school grades, which is the second step. We suppose that, to start with, we have for each student an equated college grade, to be denoted by c^*_{ijk}, which is such that a c^*_{ijk} for one college is comparable to a c^*_{ijk} from any other college. Thus, if we use the technique for equating college grades which was described in the first part of this section, then the c^*_{ijk}'s would be calculated by the formula

$$(5.13) \qquad c^{*}_{ijk} = \hat{\alpha}_j + \hat{\beta}_j \, c_{ijk} \quad .$$

Note that c^{*}_{ijk} (5.13) is not exactly the same as c_{ijk} (2.2); the latter is based on the true (unknown) values of α_j and β_j, while the former is based on estimates of α_j and β_j. We may assume a model in which c_{ijk} has conditional expectation (given T and H) of the form

$$\begin{aligned} (5.14) \qquad E(c_{ijk}) &= a' + b' h_{ijk} + b\, T_{ijk} \\ &= a' + b'(a'_i + b'_i H_{ijk}) + b\, T_{ijk} \\ &= a_i + b_i H_{ijk} + b\, T_{ijk} \end{aligned}$$

and unknown variance independent of (i,j,k). In (5.14) we are writing $h_{ijk} = a'_i + b'_i H_{ijk}$ [essentially the same thing as (2.3)], and we define $a_i = a' + b' a'_i$ and $b_i = b' b'_i$. Although the estimation of the a_i's, the b_i's, and b under the model (5.14) involves nothing more than the application of standard least-squares theory, the solution of the normal equations is a bit tricky; therefore we are treating the matter in some detail, but in the Appendix[4].

The material in the Appendix[4] is developed in terms of the c_{ijk}'s rather than the c^{*}_{ijk}'s, but in reality it is of course the c^{*}_{ijk}'s (5.13) which will have to be used in all the calculations. The assumption is made that the c^{*}_{ijk}'s are reasonably close to the c_{ijk}'s, i.e., that the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s are reasonably close to the α_j's and β_j's; by "reasonably close" is meant, roughly speaking, that the discrepancies (errors) here are small relative to other pertinent errors in the prediction system. The assumption should not be too unrealistic, in view of the fact that relatively large numbers of students from each college (as indicated by the $N_{\cdot j}$'s) would presumably have been utilized for estimating the α_j's and β_j's (note that

the $N_{\cdot j}$'s will tend to be quite large in comparison with the $N_{i\cdot}$'s).

This section, while explaining the mechanics of the first approach and the assumptions upon which it rests, has essentially not attempted any critical evaluation. The latter will be the task of Section 6.

6. Evaluation of the first approach. The first step in the first approach involves equating of the college grades by means of the C|T model. Unfortunately, though, the assumptions required by this C|T model will probably fail to be satisfied. Mainly for this reason, it appears that the first approach should not be recommended for use. Nevertheless, certain specialized situations may arise for which the first approach can be properly employed.

If all colleges select their students on the basis of T and T alone, then the assumptions of the C|T model would presumably be fully satisfied. In reality, though, the selections of most colleges would be heavily influenced by both H and T; furthermore, admissions decisions are likely to lean relatively more heavily on H in the future than at present, if a central prediction system becomes available whereby grades from different high schools can be made comparable with each other. The distortion and bias which the C|T model leads to, when there is selection on the basis of H as well as T, is perhaps best demonstrated by some illustrations.

Consider two colleges, j and J. To make matters simple, suppose that their grading standards are actually the same, so that $\alpha_j = \alpha_J$ and $\beta_j = \beta_J$.

(a) If both colleges select their students on the basis of T alone, then the C|T model should be appropriate. Thus there should essentially be no distortion or bias in the estimation of the α's and β's, and no built-in tendency for $\hat{\alpha}_j$ to exceed $\hat{\alpha}_J$ or vice versa, or for $\hat{\beta}_j$ to exceed $\hat{\beta}_J$ or

vice versa.

(b) Suppose now that college j selects on the basis of T alone, while J utilizes both H and T for selection. More specifically, suppose that j accepts each student whose T-score is above a certain minimum, while J accepts each student for whom a certain linear combination of T and H [actually, something like (2.3) should be used for H] is above a certain minimum. Since j selects on T alone, $\hat{\alpha}_j$ and $\hat{\beta}_j$ will not be distorted. But consider now what happens with college J. For each fixed value of T, the students in J will tend to have higher H-scores than those in j, since college J has a minimum H-score for each value of T whereas college J does not. Thus, since the students in J tend to have higher H-scores for a fixed value of T than the students in j, it follows that the students in J will also tend to have higher C-scores for a fixed value of T(remember that we are assuming that the grading standards are identical at the two colleges). Under the C|T model, this condition will of course force us to the false conclusion that students of equal ability receive higher grades at J than at j, whereas in fact they receive the same grades. The trouble lies in the fact that students with a given T-score at college J do <u>not</u> have the same average ability as students with the very same T-score at college j. Rather they have a higher ability (thanks to the superior admissions policy of college J), and this higher ability is reflected in higher grades. But unfortunately, under the C|T model, these higher grades are mis-interpreted as indicating lower grading standards. Thus the C|T model leads to a distorted estimate of α_J (essentially the estimate tends to be too low), and perhaps also to a distorted estimate of β_J.

(c) Consider next a situation where j and J both select on the basis of both T and H. Suppose that j accepts each student for whom a

certain linear combination of T and H is above a certain minimum, and suppose that J accepts each student for whom the very same linear combination of T and H is above a certain minimum, but suppose that J uses a higher minimum than j. Even in this case, students with a given T-score at J will tend to have higher H-scores, and hence higher C-scores, than students with the very same T-score at j. Thus it is clear that, just as in (b) above, the C|T model will lead to the false conclusion that grading standards are tougher at j than at J. This distortion will be reflected in the estimates of the equating parameters.

(d) Suppose that j and J both select on the basis of H alone; in other words, T is not used for selection at all. Suppose that J uses a minimum H-score which is higher than that used by j. Then, here again, it is easy to see that the C|T model leads us to the very same kind of distortion which troubled us in (b) and (c) above.

If the first part of the first approach ends up by giving us distorted values to use for the α_j's and β_j's in (5.13), then this distortion will certainly tend to be carried over into the estimation of the a_i's and b_i's in the second part of the first approach, although its effect will probably be diluted. But, in general, if α_j and β_j are badly distorted for a particular college j, this will tend to exert a strong distortive influence on the a_i's and b_i's of those high schools which send a proportionately large number of students to college j.

Thus the built-in bias which is evidently present in the first approach would seem to be sufficient reason for recommending against the use of this approach. However, it is always possible that, in practice, this bias might be demonstrated to be of a small enough magnitude that it would not be considered serious; but if such be the case, the burden of proof, for safety's

sake, should probably rest on those who feel that the bias is too small to be important.

One way of assessing the potential importance of this bias would be to examine the sample distribution of H given T for different colleges. In order to do this, though, one would either have to work with equated H-values, or else deal with students from a single high school at a time. The linear regression of H on T (representing the estimated mean value of H given T as a linear function of T) could be computed for each college; if these linear regressions show significant differences among colleges, then this should constitute adequate grounds for avoiding the use of the first approach. But if no such differences appear, then one might consider utilizing the first approach after all.

We now mention a second but less important drawback with the first approach. When the college equating parameters are estimated under the $C \mid T$ model in the first step of the first approach, only the information on T (and C) is utilized, and not the information on H. It would appear that the estimates of the α_j's and β_j's would be more efficient (i.e., would have smaller variance) if the information on both T and H were utilized, as is done in the third approach where the model is based on the conditional distribution of C given T and H. If the first step of the first approach results in estimates of the α_j's and β_j's which are not quite as efficient as they might be, then this loss of efficiency (probably relatively small) would be expected to carry over into the estimation of the a_i's and b_i's in the second step.

One reason for presenting the formulas and techniques of the first approach in Section 5 was that conditions might sometimes exist under which the first approach would be valid and could be applied in its entirety. A

second reason, however, was the fact that the separate parts of Section 5 will find applications in several other contexts. We shall see that formulas from the first part of the first approach will form the basis of the second part of the second approach. Also, an important use of the material of the first part of the first approach will be to provide an approximate solution (to serve as a starting point) for a system of equations which will arise in connection with the third approach and which will have to be solved by iterative procedures. Note finally that the second part of the first approach (by itself) is what would be used in the case of a single college which wants to utilize the data of its own students for predicting C on the basis of both T and H.

7. The second approach: equating the high school grades by means of the T variable, followed by equating of the college grades. The first step of the second approach is to equate the high school grades using the T-scores as the equating variable. In the case of the first step of the first approach, it was evident that the C|T model was the most reasonable of the possible models, inasmuch as there was clearly selection on the basis of T but not C. In the present case, however, the situation is not at all clear-cut, since it is not so evident what role the selection is playing. One might consider (i) a model in which the conditional distribution of H given T is normal (to be called the H|T model); (ii) a model which specifies that the joint distribution of T and H is bivariate normal (to be called the T, H model); or (iii) a model in which the conditional distribution of T given H is normal (to be called the T|H model).

If the H|T model is employed, then we use exactly the same procedure for estimating the equating parameters as in the first step of the first

approach (see Section 5); we need only replace C by H wherever C appears, and alter the subscripts. However, the H|T model may not be too appealing; this would particularly be the case if all individuals who took the test are entered into the calculations, thereby avoiding any apparent selection on the basis of T.

If the T, H model is valid, then that automatically means that both the H|T model and the T|H model are valid, since a normal joint distribution implies that all conditional distributions are normal. Hence the T, H model makes more assumptions than either the H|T model or the T|H model, and this might often be considered a disadvantage. Nevertheless, we shall present some formulas for the T, H model later in this section, after we consider the T|H model.

It might be argued that the T|H model is the most reasonable of the three, since the high school students who end up taking the test (T) may have been, at least to some extent, selected on the basis of their grades (H). Such selection could occur both through self-selection (i.e., students with better grades would feel more optimistic about college and would be more likely to register for the College Board examinations) and through the influence of teachers and counsellors, who would be more likely to encourage students with higher grades to try to go to college and to register for the test. If the joint distribution of T and H for the entire (unselected) population of high school students would be bivariate normal if they were to take the test, then it follows that the conditional distribution of T given H for the selected group (i.e., the group that actually takes the test) will be normal (with the conditional mean of T being equal to a linear function of H), provided that the selection is on the basis of H alone.

The argument in favor of this T|H model is much more convincing if the data from all students who took the test, regardless of what they ended up doing about college, is utilized in theequating. If only those students who go to college, and to a college within the system, are utilized, so that all students with T-scores (and H-scores) but no C-scores are omitted from the data, then there will almost certainly be selection on the basis of T as well as H. This would call into question the validity of the T|H model, as well as of the H|T and T, H models.

We now consider the estimation of the equating parameters under the T|H model. Under this model, the conditional distribution of T_{ijk} given h_{ijk} is of the form

$$(7.1) \qquad (2\pi\sigma^2)^{-\frac{1}{2}} \; e^{-\frac{1}{2}(T_{ijk}-a'-b'h_{ijk})^2/\sigma^2} \; .$$

Alternatively, instead of specifying our model by the distribution (7.1), we can simply write the model expectation equation

$$(7.2) \qquad E(T_{ijk}) = a' + b' \, h_{ijk} \; ,$$

which is in the form customarily used for analysis of variance problems involving linear models. Since the H_{ijk}'s and not the h_{ijk}'s are what is observed, we substitute

$$(7.3) \qquad h_{ijk} = a_i' + b_i' \, H_{ijk}$$

into (7.2) and obtain

$$(7.4) \qquad \begin{aligned} E(T_{ijk}) &= a' + b' \, (a_i' + b_i' \, H_{ijk}) \\ &= a_i + b_i \, H_{ijk} \; , \end{aligned}$$

where $a_i = a' + b'a_i'$ and $b_i = b'b_i'$. Thus the estimation of the equating parameters a_i and b_i is of course nothing but an elementary problem in analysis of variance, for which the well-known solution is given by

$$\hat{a}_i = \bar{T}_{i.} - \hat{b}_i \bar{H}_{i.} \tag{7.5}$$

and

$$\hat{b}_i = S_{THi.}/S_{HHi.} \quad . \tag{7.6}$$

(Obviously $S_{HHi.}$ must be > 0 for all i; any high school with $S_{HHi.} = 0$ would have to be thrown out.)

In case the model is made simpler by considering all b_i''s to be equal to 1 in (7.3) and (7.4), then (7.4) would become

$$E(T_{ijk}) = a_i + b'H_{ijk} \quad . \tag{7.7}$$

Under the model (7.7), the equating parameters are of course estimated by

$$\hat{a}_i = \bar{T}_{i.} - \hat{b}'\bar{H}_{i.} \quad , \tag{7.8}$$

where

$$\hat{b}' = \sum_i S_{THi.} / \sum_i S_{HHi.} \quad . \tag{7.9}$$

If there are two or more T-variables instead of just one, then the estimates of the equating parameters under the T|H model become distinctly more complicated than the simple formulas (7.5-7.6). This case is covered in the Appendix[5].

We turn now to the T, H model. Although this model may not find frequent use, we include it here for the sake of completeness. The model might be appropriate if a situation should arise where there is no selection on the basis of either T or H; such a situation might occur, e.g., if a test T is given to every student in all the schools and then the T and

H data from every student is entered into the calculations. We will restrict our consideration of the T, H model mainly to the case where there is only one T-variable, and will not attempt to treat the case of two or more T-variables except for the simplified model in which the b_i's are dropped.

The T, H model assumes a bivariate normal joint distribution for T_{ijk} and h_{ijk}. Let μ and $(1/\theta)^2$ denote respectively the mean and variance of T_{ijk}, and let ρ be the correlation coefficient between T_{ijk} and h_{ijk}. We can arbitrarily specify that $h_{ijk}(2.3)$ has mean 0 and variance 1. Then the joint distribution of T_{ijk} and h_{ijk} is given by

$$(7.10) \qquad (2\pi)^{-1} \left|\begin{matrix} 1 & \rho \\ \rho & 1 \end{matrix}\right|^{-\frac{1}{2}} \theta \exp\left[-\tfrac{1}{2}(\theta T_{ijk}-\theta\mu, h_{ijk}) \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{pmatrix} \theta T_{ijk}-\theta\mu \\ h_{ijk} \end{pmatrix}\right] .$$

Upon applying the transformation (2.3) to (7.10), we find that the joint distribution of T_{ijk} and H_{ijk} is given by

$$(7.11) \qquad (2\pi)^{-1} \left|\begin{matrix} 1 & \rho \\ \rho & 1 \end{matrix}\right|^{-\frac{1}{2}} \theta |b_i| \; e^{-\frac{1}{2}(\theta T_{ijk}-\theta\mu, a_i+b_iH_{ijk}) \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{pmatrix} \theta T_{ijk}-\theta\mu \\ a_i+b_iH_{ijk} \end{pmatrix}} .$$

In the Appendix[6] it is shown that the maximum-likelihood estimates of the equating parameters a_i and b_i under the T,H model (7.11) are given by

$$(7.12) \qquad \hat{a}_i = \hat{\rho}\,\hat{\theta}\,(\bar{T}_{i.}-\bar{T}_{..})-\hat{b}_i\bar{H}_{i.}$$

and

$$(7.13) \qquad \hat{b}_i = \frac{\hat{\rho}\,\hat{\theta}\,S_{THi.}}{2S_{HHi.}} \left[1 + \left(1 + \frac{4N_{i.}(1-\hat{\rho}^2)S_{HHi.}}{\hat{\rho}^2\hat{\theta}^2S^2_{THi.}}\right)^{\frac{1}{2}}\right] ,$$

where $\hat{\theta}$ and $\hat{\rho}$ are the values of θ and ρ which maximize the expression

$$(7.14)\quad L(\theta,\rho) = -\tfrac{1}{2} N..\log(1-\rho^2)+N..\log\theta + \sum_i N_{i.}\log\left|\frac{\rho\theta\, S_{THi.}}{2S_{HHi.}}\left[1+\left(1+\frac{4N_{i.}(1-\rho^2)S_{HHi.}}{\rho^2\theta^2 S^2_{THi.}}\right)^{\frac{1}{2}}\right]\right| - \tfrac{1}{2}\theta^2 S_{TT} - \frac{\rho^2\theta^2}{2(1-\rho^2)}\sum_i S_{TTi.} + \frac{\rho^2\theta^2}{4(1-\rho^2)}\sum_i \frac{S^2_{THi.}}{S_{HHi.}}\left[1+\left(1+\frac{4N_{i.}(1-\rho^2)S_{HHi.}}{\rho^2\theta^2 S^2_{THi.}}\right)^{\frac{1}{2}}\right] .$$

We will make no attempt to explore in detail the problem of how to find the maximum of this complicated function $L(\theta,\rho)$ (7.14), but the problem might be attacked by techniques of numerical analysis, such as the method of steepest descent.

In case θ and ρ are considered to be known, then the known values of θ and ρ should be used in (7.12-7.13) and of course the problem of maximizing (7.14) would no longer exist. The maximization of (7.14) might also be circumvented via other avenues. For example, the simple and obvious formula $\theta = (N../S_{TT})^{\frac{1}{2}}$ would probably yield a value of θ whose difference from the exact maximizing value $\hat{\theta}$ would be negligible for large $N..$. If this value of θ were substituted into (7.14), then it would remain only to maximize (7.14) with respect to the single variable ρ. Again, it might be possible to find a relatively straightforward formula which, for large $N..$, would yield a value of ρ that would be only negligibly different from the one ($\hat{\rho}$) which gives the exact maximum of (7.14).

We consider now the simpler but more restrictive T,H model in which all the b_i's in (2.3) are set equal to 1. For our purposes, the joint distribution of T_{ijk} and H_{ijk} under this model is most conveniently written in the form

$$(7.15) \qquad (2\pi)^{-1} |\Sigma|^{-\frac{1}{2}} \; e^{-\frac{1}{2}(T_{ijk}-\mu,\, a_i+H_{ijk})\Sigma^{-1}\begin{pmatrix} T_{ijk}-\mu \\ a_i+H_{ijk}\end{pmatrix}} \quad ,$$

where Σ is the 2x2 variance matrix. We show in the Appendix[7] that the maximum-likelihood estimate of a_i under this model (7.15) is

$$(7.16) \qquad \hat{a}_i = -\bar{H}_{i.} + \frac{\sum_{i'} S_{THi'.}}{\sum_{i'} S_{TTi'.}} \left(\bar{T}_{i.} - \bar{T}_{..}\right) \quad .$$

Thus the estimation of the equating parameters turns out to be much easier under the model (7.15) than under the model (7.11). If the model (7.15) is generalized to allow for two or more T-variables, then the estimation of the a_i's still presents little difficulty, although the inversion of a matrix is required[7].

This completes our discussion of the first step (equating of high school grades via the T-scores) of the second approach. We now turn to the second step, which consists of the equating of the college grades on the basis of the T-scores and the equated H-scores. We suppose that, at the start of the second step, we have for each student an equated high school grade, to be denoted by h^*_{ijk}, which is such that an h^*_{ijk} for one high school is comparable to an h^*_{ijk} from any other high school. Thus the h^*_{ijk}'s would be calculated by a formula of the form

$$(7.17) \qquad h^*_{ijk} = \hat{a}_i + \hat{b}_i H_{ijk}$$

if any of the methods presented earlier in this section are utilized for equating the high school grades. Now h^*_{ijk} (7.17) is not quite the same thing as h_{ijk}(2.3), since the latter is based on the exact but unknown values of the equating parameters. In what follows, we shall present our development in

terms of the h_{ijk}'s in order to keep everything rigorous. However, for practical purposes the unknown h_{ijk}'s could not, of course, be used, and the h^*_{ijk}'s would have to be used instead with the assumption that they would be reasonably close to the h_{ijk}'s.

We may assume a model in which the conditional distribution of the equated college grade (2.2) given T_{ijk} and h_{ijk} is normal with variance arbitrarily taken to be equal to 1. Then this conditional distribution, when expressed in terms of the observed quantity C_{ijk} (rather than in terms of c_{ijk}), is of the form

$$(7.18) \qquad (2\pi)^{-\frac{1}{2}} |\beta_j| \, e^{-\frac{1}{2}(\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-\nu^\circ h_{ijk})^2} \quad .$$

Note, in fact, that this model is formally identical with the one covered in Note 3 of the Appendix, and that (7.18) is the same thing as (A3.1) except that (7.18) has h_{ijk} where (A3.1) uses the notation T°_{ijk}. Thus it turns out that the problem of estimating the α_j's and β_j's in the second step of the second approach is essentially the same problem as one which was previously encountered and dealt with in Section 5 in connection with the first step of the first approach. Because of this, we need not consider the problem further, except to point out again its solution: the estimates of the α_j's and β_j's are calculated by formulas (A3.7) and (A3.9) respectively, but only after the rather complicated system (A3.10 - A3.11) has been solved for ν and ν°. In case there is more than just the one T-variable represented in (7.18) [i.e., more than just the two T-variables represented in (A3.1)], then the theory of Note 3 generalizes in a straightforward manner.

It might be argued that the estimation of the α_j's and β_j's in the second step of the second approach could be omitted altogether. If it is

desired to predict only the college grades on the common measuring scale (the c_{ijk}'s) rather than the grades at individual colleges (the C_{ijk}'s), then for this purpose there would be no need to estimate the α_j's and β_j's. A college could compare its applicants with each other on the basis of their predicted c_{ijk}'s just as effectively as on the basis of their predicted C_{ijk}'s for that college, since one is just a linear function of the other. On the other hand, though, the college might like to know the relationship between the c_{ijk}'s and its own grades (C_{ijk}'s); in other words, it might like to find out what its own $\hat{\alpha}_j$ and $\hat{\beta}_j$ are (and it might also, incidentally, be curious about the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s of other colleges). We might also note that, even for the purposes of predicting just the c_{ijk}'s, it would still be necessary to obtain estimates of ν and ν°, so that we would apparently have to solve the system (A3.10-A3.11) to get $\hat{\nu}$ and $\hat{\nu}^\circ$ in any event.

8. Evaluation of the second approach. The first step of the second approach involves estimation of the a_i's and b_i's, the equating parameters for the high school grades, by means of the H|T model, the T, H model, or the T|H model. In Section 7 we indicated doubt that the assumptions underlying any of these three models would be satisfied so long as the calculations in the first step were based only on those students for whom C-scores (as well as T-scores and H-scores) were available, since such a group of students would probably have been selected on the basis of both T and H. On the other hand, it was suggested that the T|H model might well be valid if all students who took the test were included in the calculations, since it could then be argued that there was selection on H but not on T; even then, however, there is still some possibility of selection on T, as we shall see below. Finally, we will want to consider the question of how the $\hat{a}_i$'s

and $\hat{b}_i$'s of the second approach compare in efficiency with the corresponding estimators under the third approach, but we shall defer this matter to Section 10.

Our evaluation of the second approach is based on the factors just mentioned. Thus it appears that the second approach should not be recommended at all if the data used in the first step is restricted to those students for whom C-scores are reported. But if the data from all students who took the test is used in the first step, and if the calculations are based on the T|H model, then the results of the second approach may be reasonably satisfactory. Even so, it appears that this special form of the second approach (i.e., using the T|H model and including all students who took the test in the first step) still rests on less solid assumptions than the third approach, and therefore should not be preferred over the third approach unless we are trying to avoid the latter because of some reason such as excessive calculation costs. On the other hand, the variance of the estimators of the a_i's and b_i's apparently can easily favor our special form of the second approach, depending on conditions (see Section 10). But, as we shall see later in this section, it is possible for a certain type of systematic bias to creep in if we use the special form of the second approach; since this type of bias cannot arise with the third approach, the third approach therefore appears safer.

We now turn to a more detailed explanation of a couple of the points which were summarized in the first paragraph of this section. We first consider what happens if only those students with C-scores are utilized in the first step of the second approach, so that, of those students who took the test (T), the ones are excluded who either failed to go to college at all or else went to a college outside the system. Now it is rather safe to assume that the students who failed to go to college at all received relatively low

T-scores and H-scores. The students who did go to college, but to a college outside the system, would also have relatively low T-scores and H-scores if the colleges outside the system tend to be lower-quality institutions than the colleges inside the system. Thus it is logical to expect that there will be some selection on the basis of both T and H with respect to the determining of whether or not a student becomes part of the group receiving C-scores. Now note that, if the unselected group has a bivariate normal distribution, then neither the H|T model nor the T,H model will be valid if there is selection on H, and neither the T|H model nor the T,H model will be valid if there is selection on T; thus none of the three models will be valid if there is selection on both H and T. In fact, in the idealized situation where selection occurs strictly on the basis of whether a certain linear combination of a student's T-score and H-score exceeds a certain minimum, the average value of H given T will no longer even be a linear function of T, and the average value of T given H will no longer be a linear function of H. It would appear that distortion in the estimation of the a_i's and b_i's would be particularly marked with respect to the relation between the $\hat{a}_i$'s and $\hat{b}_i$'s of, say, two high schools which differed substantially in their average T-scores and H-scores, or which might differ with respect to the influence of the selection for the different values of T and H. Consider the following special example. Suppose that two high schools i and I differ substantially in their average H-scores, but that they have exactly the same grading standards (i.e., $a_i = a_I$ and $b_i = b_I$). Suppose that the selection is on exactly the same basis for both schools. If we use the T|H model, then the estimation process will essentially try to estimate, for each high school, the average value of T given H, as a linear function of H. But the true average value of T given H will be a curvilinear function of H, since there is

selection on both T and H. Thus the estimation for schools i and I will essentially produce two linear approximations of the same curvilinear function. However, these two linear approximations will tend to be rather different from each other, since the H data from the two schools will be concentrated on different parts of the H-axis, for which the best linear approximations of the curvilinear function would of course be different. Thus we could easily end up with $(\hat{a}_i, \hat{b}_i)$ radically different from $(\hat{a}_I, \hat{b}_I)$, when in fact they should be almost the same since the two grading standards are identical.

We consider next what happens if, in the first step of the second approach, all students who took the test (T) are entered into the calculations, rather than just the students for whom C-scores were reported. Such a scheme might or might not produce additional administrative problems, but in any event it should result in substantially less distortion in the $\hat{a}_i$'s and $\hat{b}_i$'s when the second approach is used (with the T|H model). If all students who took the test are included in the calculations, then it would appear, superficially, at least, that there was no selection on the basis of T. Now there would still seem to be selection on the basis of H, for reasons previously mentioned in the earlier part of Section 7: not all high school students sign up to take the test, and those who do are probably partially selected on the basis of H. If there is selection on H, then both the H| T model and the T, H model are invalid;but the T|H model is still valid, so long as no selection on T has crept in.

We might feel that it would be reasonable to assume that there is no selection on T, so that the T|H model would be fully valid. Nevertheless, it would perhaps be best to try to examine this assumption closely. We shall now try to present an argument which advances the point of view that there

might indeed be selection on T, and that such selection could distort the $\hat{a}_i$'s and $\hat{b}_i$'s. Suppose that high school students are able to form some rough idea of how well they would do on the test (T) if they were to take it. For example, consider a bright individual who has loafed all through high school and received grades (H) which are poor in relation to his ability. Such an individual might easily recognize that he would score quite well on the test in comparison with other persons having the same H-score as his, and, for this reason, he and others like him might be more likely to register for the test than individuals who would expect to have average or below-average T-scores in relation to their H-scores. Thus the conditional distribution of T given H would be different in the selected population (those who sign up for the test) than in the unselected population (all high school students). If the latter distribution were normal (with the mean being a linear function of H), then the former distribution would not be, except perhaps by sheerest coincidence. Thus selection on the basis of T would be present, and would render invalid the T|H model. (Whether this kind of selection is really occurring, incidentally, might be checked by an experiment. After the test (T) has been administered to those who register for it, it could be administered again, necessarily free of charge, to all students in a few selected high schools who did not take it previously. The distribution of T given H for the latter group could then be compared with the distribution of T given H for the former group, to see if a difference really existed.) To see how a serious distortion in the $\hat{a}_i$'s and $\hat{b}_i$'s could possibly occur, consider the following situation. Suppose we have two high schools, i and I, such that i is in a low-income area and I is in a high-income area. Suppose that i and I have the same grading standards (i.e., $a_i=a_I$ and $b_i=b_I$.) Now it would not be surprising if, for any given H-score, more students in I than in i register for the test, inas-

much as students who can better afford college would probably be more likely to try to go to college. Furthermore, the relatively few students in i who do register for the test might be mainly the ones who would be capable of getting such high T-scores (in relation to their H-scores) that they would be awarded scholarships, and could thereby afford to go to college. In I, on the other hand, it would be not merely the potential scholarship awardees who would register for the test, but also a large number of students of lesser ability (i.e., with lower anticipated T-scores in relation to their H-scores) who could afford to go to college even without a scholarship. Thus, among the individuals registering for the test, the average T-score for any given H-score would be higher in i than in I, due to the differential selection involving both T and H. Hence $\hat{a}_i$ would generally tend to be higher than $\hat{a}_I$, thereby indicating (falsely) that i has tougher grading standards than I. It cannot be foretold with any certainty whether this type of distortion is likely to occur in actual practice, or whether it is merely a slim theoretical possibility; unfortunately, though, a bias which has an economic or sociological basis is probably one of the most undesirable types of biases that could be built into a prediction system of this kind, and so the possibility of such a bias would evidently have to be carefully investigated and ruled out before the bias could safely be assumed not to exist.

In closing this section, we point out some potential difficulties in connection with the second step of the second approach. In the first place, it should be apparent that, if systematic biases creep into the $\hat{a}_i$'s and $\hat{b}_i$'s in one way or another during the first step of the second approach, then these biases will be carried over into the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s when the latter are determined in the second step of the second approach. Although the effect of the distortion in the $\hat{a}_i$'s and $\hat{b}_i$'s may be somewhat diluted by the time it

reaches the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s, it may still make itself felt rather strongly on the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s of those colleges which draw a proportionately large number of students from high schools whose $\hat{a}_i$'s and $\hat{b}_i$'s were badly distorted.

A second difficulty in connection with the second step of the second approach becomes evident when we note that homoscedasticity (equality of variance) is assumed in the model (7.18). Actually, it is really the h^*_{ijk}'s (7.17) rather than the h_{ijk}'s which are utilized in the calculations. In general, h^*_{ijk} will tend to be closer to h_{ijk} the larger $N_{i.}$ is and the smaller $|H_{ijk} - \overline{H}_{i.}|$ is. Thus the conditional variance of c_{ijk} given T_{ijk} and h^*_{ijk} presumably would be smaller the larger $N_{i.}$ is and the smaller $|H_{ijk} - \overline{H}_{i.}|$ is. This means that the homoscedasticity assumption is not strictly satisfied, since the conditional variance is not the same for all (i,j,k). But the resulting effect on the estimates of the α_j's and β_j's and of ν and ν° would probably not be too serious, because of the large number of students upon whom these estimates are based. However, it seems desirable to at least call attention to this difficulty, even though it appears to be only a minor one. There will apparently also be other minor difficulties related to the fact that the conditional distribution of C_{ijk} given T_{ijk} and h_{ijk} (7.18) is not exactly the same thing as the conditional distribution of C_{ijk} given T_{ijk} and h^*_{ijk}.

9. <u>The third approach: equating the high school grades and equating the college grades simultaneously</u>. In this third approach, the high school equating parameters and the college equating parameters are estimated simultaneously in a single step, rather than in two separate steps as was the case with each of the first two approaches. We assume a model in which the conditional distribution of c_{ijk} given T_{ijk} snd h_{ijk} is normal with variance equal to 1 and with mean equal to a linear function of T_{ijk} and h_{ijk}, so that

the distribution can be written in the form

$$(9.1) \qquad (2\pi)^{-\frac{1}{2}} \; e^{-\frac{1}{2}(c_{ijk}-\mu-\nu T_{ijk}-b'h_{ijk})^2} \; .$$

The observed data, though, is in terms of the C_{ijk}'s and H_{ijk}'s rather than the c_{ijk}'s and h_{ijk}'s. In (9.1) we thus apply the transformation of variable (5.2-5.3) and the substitution (7.3) to obtain

$$(9.2) \qquad (2\pi)^{-\frac{1}{2}} |\beta_j| \; e^{-\frac{1}{2}(\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-a_i-b_i H_{ijk})^2}$$

as the conditional distribution of C_{ijk} given T_{ijk} and H_{ijk}. In (9.2) we are defining $\alpha_j = \alpha'_j - \mu$, $a_i = b'a'_i$, and $b_i = b'b'_i$, where α'_j, a'_i, and b'_i are as indicated in (5.2) and (7.3). Note that the parameter μ could just as well be absorbed in a_i as in α_j; this reflects a certain indeterminacy which is present.

A model of the form (9.2) we will call the $C|T,H$ model. If the joint distribution of C,T, and H in the unselected population (of all high school students) is trivariate normal, and if selection is made in any fashion whatever on the basis of T and H and of T and H alone, then the conditional distribution of C given T and H will be normal (with the mean a linear function of T and H) for the selected population (i.e., individuals for whom C-scores are available) as well as for the unselected population. This is the reason for utilizing the $C|T$, H model.

The calculations for obtaining the estimates of the equating parameters under the $C|T$, H model (9.2) are somewhat involved. We present here all the basic formulas, and in the Appendix[8] we give their derivation.

Let us recollect the notation

(9.3) $\quad d_{ij} = H_{ij} - N_{ij}\bar{H}_{i.}, \quad e_{ij} = T_{ij} - N_{ij}\bar{T}_{i.}$.

The estimates of the high school equating parameters are given by

$$(9.4) \qquad \hat{a}_i = (1/N_{i.}) \left(\sum_j N_{ij}\hat{\alpha}_j + \sum_j \hat{\beta}_j C_{ij}\right) - \hat{\nu}\bar{T}_{i.} - \hat{b}_i\bar{H}_{i.}$$

and

$$(9.5) \qquad \hat{b}_i = (1/S_{HHi.}) \left[\sum_j d_{ij}\hat{\alpha}_j + \sum_j \hat{\beta}_j(S_{CHij} + \bar{C}_{ij}d_{ij}) - \hat{\nu}\, S_{THi.}\right] ,$$

where $\hat{\nu}$, the $\hat{\alpha}_j$'s, and the $\hat{\beta}_j$'s are determined by the means outlined below.

At this point we need to define some matrices. Let $G([n+1] \times n)$ be a matrix whose general element in the j-th row and J-th column is

$$(9.6) \qquad g_{jJ} = -\delta_{jJ}C_{.j} + \sum_i \frac{N_{ij}C_{iJ}}{N_{i.}} + \sum_i \frac{d_{ij}(S_{CHiJ} + \bar{C}_{iJ}d_{iJ})}{S_{HHi.}}$$

for the first n rows, and

$$(9.7) \qquad g_{\nu J} = \sum_i (S_{CTiJ} + \bar{C}_{iJ}e_{iJ}) - \sum_i \frac{S_{THi.}(S_{CHiJ} + \bar{C}_{iJ}d_{iJ})}{S_{HHi.}}$$

for the $(n+1)$-th row, where we use the subscript ν rather than $(n+1)$ to refer to the bottom row. In (9.6), δ_{jJ} is the Kronecker delta; that is, $\delta_{jJ} = 0$ if $j \neq J$ and $\delta_{jJ} = 1$ if $j = J$. Next we introduce a symmetric $([n+1]\times[n+1])$ matrix

$$(9.8) \qquad F = \left[\begin{array}{ccc|c} f_{11} & f_{12} \cdots & f_{1n} & f_{1\nu} \\ f_{21} & f_{22} \cdots & f_{2n} & f_{2\nu} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} \cdots & f_{nn} & f_{n\nu} \\ \hline f_{\nu 1} & f_{\nu 2} \cdots & f_{\nu n} & f_{\nu\nu} \end{array}\right]$$

whose elements are defined by the equations

$$(9.9) \qquad f_{jJ} = \delta_{jJ} N_{.j} - \sum_i \frac{N_{ij} N_{iJ}}{N_{i.}} - \sum_i \frac{d_{ij} d_{iJ}}{S_{HHi.}} \quad ,$$

$$(9.10) \qquad f_{j\nu} = -T_{.j} + \sum_i N_{ij} \bar{T}_{i.} + \sum_i \frac{d_{ij} S_{THi.}}{S_{HHi.}} \quad (= f_{\nu j}) \ ,$$

and

$$(9.11) \qquad f_{\nu\nu} = \sum_i \left(S_{TTi.} - \frac{S^2_{THi.}}{S_{HHi.}} \right) \quad .$$

The formulas for the $\hat{\alpha}_j$'s and $\hat{\nu}$ will be in terms of the $\hat{\beta}_j$'s: we solve the equation system

$$(9.12) \qquad F \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \nu \end{pmatrix} = G \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

for $\alpha_1, \alpha_2, \ldots, \alpha_n$, ν . Now F is not of the full rank [see (9.14) below], and so F^{-1} does not exist. F will generally be [9] of rank n. If the matrix F* ([n+1] x [n + 1]) denotes [9] a conditional inverse of F, then a solution of (9.12) is

$$(9.13) \qquad \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_n \\ \hat{\nu} \end{pmatrix} = F^*G \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} \quad .$$

The Appendix [9] indicates a way of obtaining F*. Incidentally, the solution of (9.12) for the α_j's will not be unique; if the same constant is added to each member of a set of α_j's satisfying (9.12), then the new set of values will also constitute a solution, inasmuch as

$$(9.14) \qquad \sum_{J=1}^{n} f_{jJ} = 0 \ , \qquad \sum_{J=1}^{n} f_{\nu J} = 0 \quad .$$

To obtain the $\hat{\beta}_j$'s, we have to solve a non-linear equation system. We define a matrix $U(n \times n)$ whose general element is

$$(9.15) \qquad u_{jJ} = \delta_{jJ} \sum_i \sum_k C_{ijk}^2 - \sum_i \frac{C_{ij}C_{iJ}}{N_{i.}} - \sum_i \frac{(S_{CHij}+\bar{C}_{ij}d_{ij})(S_{CHiJ}+\bar{C}_{iJ}d_{iJ})}{S_{HHi.}} \quad .$$

Then we define

$$(9.16) \qquad A(n \times n) = U - G'F^*G \quad .$$

We use $\underline{\beta}$ $(n \times 1)$ to denote the vector $(\beta_1, \beta_2, \ldots, \beta_n)'$, $\underline{N}$ $(n \times 1)$ to denote the vector $(N_{.1}, N_{.2}, \ldots, N_{.n})'$, and $D_\beta(n \times n)$ to denote a diagonal matrix whose diagonal elements are $\beta_1, \beta_2, \ldots, \beta_n$. The $\hat{\beta}_j$'s are found by solving the system

$$(9.17a) \qquad A\,\underline{\beta} = D_\beta^{-1}\underline{N} \quad .$$

for $\underline{\beta}$. The system (9.17a) can be written alternatively in the form

$$(9.17b) \qquad \sum_{J=1}^{n} a_{jJ}\,\beta_J = N_{.j}/\beta_j \qquad (j = 1,2, \ldots, n) \quad ,$$

where a_{jJ} is the general element of $A(9.16)$.

This system (9.17) has no more than one solution for the β_j's such that all β_j's are positive.[10] Possible iterative techniques for solving the system (9.17) are indicated in the Appendix; the generalized Newton-Raphson method and the method of steepest descent are considered[11].

Having just presented all the details, we now bring everything together and summarize what has to be done to obtain the estimates of the equating parameters under the C|T, H model (9.2). After calculating the elements of $F(9.8 - 9.11)$, we obtain a conditional inverse F^*, which essentially involves the inversion of an $n \times n$ matrix[9]. Then we calculate $G(9.6 - 9.7)$, $U(9.15)$, and $A(9.16)$. Next we solve the system (9.17) to get the estimates

of the β_j's. These estimates are plugged into the right-hand side of (9.13) in order to obtain the $\hat{\alpha}_j$'s and $\hat{\nu}$. Finally, the $\hat{\alpha}_j$'s, the $\hat{\beta}_j$'s, and $\hat{\nu}$ are substituted into (9.4 - 9.5) in order to get the $\hat{a}_i$'s and $\hat{b}_i$'s.

Evidently the two most troublesome steps in this procedure will be to find F* and to solve the system (9.17). In assessing the magnitude of the computing problem which we will face in these two steps, we should recall that n (the number of colleges) is anticipated to be about 400 or 500.

If the C|T, H model is altered so that the b_i's of (7.3) are eliminated (i.e., set equal to 1), then the formulas for estimating the equating parameters are a bit different from the above, although the calculations will be almost as lengthy[12]. If the C|T, H model is altered so that the β_j's are eliminated (set equal to 1), then the procedure for getting the estimates will be virtually the same as the one given above except that we no longer will have the burden of solving the non-linear system (9.17)[13]. If it is desired to use more than one T-variable in connection with the C|T, H model, due to there being selection on the basis of more than one T-variable, then arrangements for the additional T-variable(s) can be incorporated into the calculation procedure with a minimum of difficulty[14].

10. Evaluation of the third approach, and comparison with the second approach. In this section, we compare the third approach with what seems to be its leading competitor, viz., that special form of the second approach in which (see Sections 7-8) the first step utilizes all students who took the test and is based on the T|H model. The third approach apparently rests on more plausible assumptions than the special form of the second approach, but at the same time requires greater computational effort. The third approach utilizes data from a smaller group of students; this factor may have both its

drawbacks and its advantages. Finally, we will need to compare the variances of the $\hat{a}_i$'s and $\hat{b}_i$'s under the third approach and the special form of the second approach. We now consider these various points in detail.

As we noted in Section 8, the assumptions underlying the special form of the second approach may be open to some question, because of the possibility of selection on T even when all students who took the test are included in the calculations of the first step. Such a danger cannot arise with the third approach, however: the third approach will not be invalidated by there being selection on T as well as H. This is because the third approach is based on a distribution (the conditional distribution of C given T and H) which holds both T and H fixed.

Although the third approach thus seems to rest on somewhat more reasonable assumptions than the special form of the second approach, the third approach is still not completely immune from conditions of selection which might cause the assumptions of the C|T, H model to be violated. For example, if there is selection based on C(which seems unlikely), this could obviously cause trouble. Again, if some colleges are using a third predictor (in addition to T and H), and if this third predictor is successful in improving the prediction of C, then the assumptions underlying the third approach would no longer hold. However, if some college(s) should indeed discover on their own such a third predictor which genuinely does improve appreciably the prediction of C, then such a third predictor could and probably would be quickly included among the predictors employed by the central prediction system. This third predictor would probably be in the form of an additional T-variable, and, as we have already seen in Note 14 of the Appendix, the introduction of an additional T-variable causes only a relatively small increase in the computational effort required for the third approach.

All in all, we conclude that the third approach appears to be a some-

what safer choice, because of the greater plausibility of the assumptions underlying it. At the same time, we have to recognize that the computational burden is much heavier for the third approach than for the special form of the second approach. For the third approach, the calculation of the various elements (9.6 - 9.7, 9.9-9.11) of G and F is no small matter in itself, but evidently the two biggest computational tasks are to invert the large matrix F_{11} (see Note 9) and to solve the system (9.17) (see Note 11). Each of these tasks involves an (n x n) matrix (F_{11} and A respectively), and we are anticipating that n, the number of colleges, will be about 400 or 500. However, in inverting F_{11}, we essentially have the job of inverting a matrix which has large positive diagonal elements and small off-diagonal elements, a condition which may greatly simplify the inversion problem; and, in solving the non-linear system (9.17), we may be able to employ a relatively quick and simple iterative technique by making use of the suggestions in Note 11. In any event, the final assessment of how great a disadvantage the heavy computational burden of the third approach is will have to be made in terms of the estimated total cost of the required computations. For some phases of the computing for the third approach, it might be more economical to purchase a small amount of time on a rather large computer. However, decisions such as these, as well as the cost estimates, would have to be left to computer specialists.

Data from fewer students is utilized in the third approach than in the first step of the special form of the second approach. From high school i, let there by $N'_{i.}$ students who took the test (T) (and for whom H-scores are assumed to be available). Let there be $N_{i.}$ out of these $N'_{i.}$ students for whom C-scores are also available; in other words, there are $(N'_{i.} - N_{i.})$ out of the $N'_{i.}$ students who either didn't get to college at all or else went to a college outside the system. Then the third approach utilizes data (on C,T,

and H) just from the $N.. = \sum_i N_{i.}$ students who received C-scores, whereas the first step of the special form of the second approach estimates the a_i's and b_i's by using T and H data from all $N'.. = \sum_i N'_{i.}$ students (although the second step of the second approach necessarily uses data only from the N.. students).

The fact that the third approach utilizes data from a smaller group of students may have both its advantages and disadvantages. In the first place, it is possible, particularly when the central prediction system is first being started up, that there might be some extra administrative problems involved in getting the H-scores of the $(N'.. - N..)$ individuals for whom no C-scores are reported. If this should be the case, then the third approach would have the advantage of avoiding these extra administrative problems and their associated costs.

On the other hand, the third approach may be at a disadvantage for certain other reasons related to the differences mentioned above. For one thing, the smaller the ratio $N_{i.}/N'_{i.}$, the less favorably the variance of $\hat{a}_i$ (or $\hat{b}_i$) under the third approach compares with the variance of $\hat{a}_i$(or $\hat{b}_i$) under the special form of the second approach, as we shall see in more detail later in this section. There is a second possible disadvantage for the third approach which relates to the variables which are used rather than to the numbers of students which are used. Since all three variables (C, T, and H) are used in estimating the a_i's and b_i's under the third approach, the estimates of the a_i's and b_i's will necessarily have to be based on data from students who already graduated from high school the previous year (or earlier), inasmuch as we have to wait until the students have received their C-scores before we can calculate the $\hat{a}_i$'s and $\hat{b}_i$'s. With the second approach, though, only the two variables T and H are used in getting the $\hat{a}_i$'s and $\hat{b}_i$'s. We may, if we wish, base our calculations of the $\hat{a}_i$'s and $\hat{b}_i$'s in the second approach on

the same class of students (probably the college freshman class) that would be used with the third approach. But, on the other hand, we also have the option of using a later (younger) class of students, inasmuch as we do not need to wait for their C-scores; this would result in more up-to-date estimates of the a_i's and b_i's, which might or might not be an important advantage depending on the extent to which the a_i's and b_i's tend to change in the course of time. [The parameters ν and ν° of (7.18), as well as the α_j's and β_j's, would of course still have to be estimated from the earlier data, however.]

At this point, we pause to mention some rather simple types of checks which can be run to determine whether a prediction system is operating as it should in certain respects, or to compare two or more different predictive techniques. These checks will be applicable to all the different approaches and methods. Let $\hat{C}_{ijk}$ denote the predicted college grade for individual (i,j,k), and let C_{ijk} denote (as always) his actual grade. The $\hat{C}_{ijk}$'s are based on the T_{ijk}'s and H_{ijk}'s, and on parameters estimated probably from the previous year's students; the C_{ijk}'s, of course, do not become available until some time after the $\hat{C}_{ijk}$'s. Some checks which can be made are as follows. We calculate

$$D_{ijk} = \hat{C}_{ijk} - C_{ijk} \tag{10.1}$$

for each student. We then group the D_{ijk}'s (10.1) according to the high schools (i), and again according to the colleges (j). We tally the number of positive and negative D_{ijk}'s within each high school, and again within each college; if there are high schools or colleges for which the ratio of positive to negative D_{ijk}'s differs radically from fifty-fifty, this may indicate trouble with respect to the $\hat{a}_i$'s and $\hat{b}_i$'s of these high schools or the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s of these colleges respectively. For identifying trouble spots in

the high school, it might also help to arrange the students of a high school in order according to their T-scores, according to their H-scores, and/or according to a linear combination of T and H, and then perform a run test on the signs of the D_{ijk}'s for any or all of these three arrangements. The D_{ijk}'s within a high school (and, to a lesser extent, within a college) will of course not be independent since the $\hat{C}_{ijk}$'s are affected by a common $\hat{a}_i$ and $\hat{b}_i$ (a common $\hat{\alpha}_j$ and $\hat{\beta}_j$ in the case of a college), but the checks suggested above may nevertheless provide some useful indications.

A different type of check can be based on the D^2_{ijk}'s. For example, we might calculate

$$(10.2) \qquad (1/N_{.j}) \sum_i \sum_k D^2_{ijk}$$

for each college, and compare the quantities (10.2) with their expectations. If there are certain colleges for which (10.2) is inordinately large, this might indicate trouble with their $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s, but there could be an alternative explanation as well: the criterion variable (C) at such colleges might be something essentially different from the criterion variable at the vast majority of colleges, as might happen, e.g., as a result of unconventional or different curriculums. A check might also be made by calculating a quantity similar to (10.2) for each high school, and making appropriate comparisons. If there are certain high schools for which the quantity is inordinately large, this might indicate trouble with their $\hat{a}_i$'s and $\hat{b}_i$'s, or it might possibly reflect poorly-controlled or non-uniform grading techniques in these high schools.

Checks such as the ones just mentioned might be very helpful in reaching a decision between the third approach and the special form of the second approach. Both approaches could be applied to the same body of data (taken,

perhaps, from a relatively small number of high schools and colleges), and the results of the above checks for the two approaches could be compared.

We turn now to a comparison of the variances of the $\hat{a}_i$'s and $\hat{b}_i$'s obtained under the third approach versus the variances of the $\hat{a}_i$'s and $\hat{b}_i$'s obtained under the special form of the second approach. Actually, what we will do will be to make the comparison on the basis of the variance of $(\hat{a}_i + \hat{b}_i H)$ rather than the variance of $\hat{a}_i$ and of $\hat{b}_i$. After necessary adjustments for the regression parameters that are absorbed in the a_i's and b_i's, the ratio of the variance of $(\hat{a}_i+\hat{b}_i H)$ under the third approach to the variance of $(\hat{a}_i + \hat{b}_i H)$ under the special form of the second approach is[15] approximately

$$(10.3)\qquad \frac{\left[\frac{1}{N_{i.}} + \frac{(H-\bar{H}_{i.})^2}{S_{HHi.}}\right]}{\left[\frac{1}{N'_{i.}} + \frac{(H-\bar{H}'_{i.})^2}{S'_{HHi.}}\right]} \quad \frac{\rho_{TH}^2(1-\rho_{CT}^2-\rho_{CH}^2-\rho_{TH}^2+2\rho_{CT}\rho_{CH}\rho_{TH})}{(\rho_{CH}-\rho_{CT}\rho_{TH})^2},$$

where ρ_{CT}, ρ_{CH}, and ρ_{TH} denote the correlation coefficients among C, T, and H in the unselected population. $S'_{HHi.}$ in (10.3) denotes the same thing as $S_{HHi.}$ except that it is figured with respect to all $N'_{i.}$ individuals rather than just over the $N_{i.}$ individuals, and $\bar{H}'_{i.}$ denotes the mean over all $N'_{i.}$ individuals. The variance formulas for the third approach which were used in obtaining (10.3) were figured on the basis of a simplified model in which the α_j's and β_j's were assumed to be known exactly[15]; for this reason, the ratio (10.3) is presumably slightly smaller than it should be, although the discrepancy is probably not great since the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s are based on such large numbers ($N_{.j}$'s) of students.

One might perhaps suspect intuitively that the third approach should produce estimators having smaller variance than those produced by the second approach, inasmuch as the former approach utilizes information from all three

variables (C, T, and H) whereas the latter utilizes only the T and H information. Thus, except for the factors enclosed by the square brackets, one might suspect that the ratio (10.3) ought to be < 1. However, it turns out that this will often not be the case. The explanation lies in the different assumptions and distributions upon which the $C|T, H$ model and the $T|H$ model are based.

The ratio of the two factors in square brackets in(10.3) should be roughly $N'_{i.}/N_{i.}$ in most cases. If these two factors are omitted, what remains of (10.3) is

$$(10.4) \qquad \rho^2_{TH}(1-\rho^2_{CT}-\rho^2_{CH}-\rho^2_{TH}+2\rho_{CT}\rho_{CH}\rho_{TH})/(\rho_{CH}-\rho_{CT}\rho_{TH})^2 \quad ,$$

which is the ratio of the conditional variance of C given T and H to the conditional variance of T given H (under a trivariate normal distribution, which is assumed[15]), multiplied by a factor which adjusts for the different regression coefficients that are absorbed in a_i and b_i under the two approaches.

It may be instructive to calculate (10.4) for a couple of numerical examples. Suppose $\rho_{CT} = .70$, $\rho_{CH} = .50$, and $\rho_{TH} = .60$. Then the value of (10.4) is 18.0. Again, suppose $\rho_{CT} = .70$, $\rho_{CH} = .65$, and $\rho_{TH} = .50$. This time (10.4) is equal to 0.81. Note how a small change in the ρ's can produce a violent change in (10.4).

Thus we see that (10.4) can easily favor the special form of the second approach rather strongly (by being much greater than 1), although it can also favor the third approach with certain values of the ρ's. The factor $(N'_{i.}/N_{i.})$ obviously will always favor the special form of the second approach. Hence, with respect to the criterion (10.3), the third approach may not show up at all favorably. On the other hand, if the special form of the second approach results in biased estimates of the a_i's and b_i's whereas the third approach

does not, then the magnitude of such biases could possibly dwarf any counteracting advantage accruing from (10.3) being > 1. Also, we will be able to see from some formulas in the next section that, insofar as the expectation of D^2_{ijk} [see (10.1)] is concerned, the contribution which the variance of $(\hat{a}_i + \hat{b}_i H)$ makes to this expectation will be relatively small even if the third approach is used when (10.3) is much larger than 1.

In closing this section, we pose the question of whether the a_i's and b_i's might be estimated by some approach which would utilize both the information from the T|H model and the information from the C|T, H model, and thereby produce $\hat{a}_i$'s and $\hat{b}_i$'s whose variances would be better than the corresponding variances under either the third approach or the special form of the second approach. We have made no attempt to explore this potentially complicated question. It is possible, though, that we might end up by running into formidable difficulties of either a theoretical or computational nature.

11. Determination of the predicted college grades, and of confidence intervals for these grades. After the estimates of all of the equating parameters and regression parameters have been calculated, it still remains to obtain predicted college grades, and perhaps confidence intervals therefor, for the current applicants for college admission. We will use $\hat{c}_{ijk}$ to denote the predicted value of the equated college grade c_{ijk} (2.2), and our presentation here will explicitly be only in terms of the $\hat{c}_{ijk}$'s rather than the $\hat{C}_{ijk}$'s. However, the latter can be obtained from the former via the formula

$$\hat{C}_{ijk} = (1/\hat{\beta}_j)(\hat{c}_{ijk} - \hat{\alpha}_j) \quad . \tag{11.1}$$

Presumably the $N_{\cdot j}$'s will be large enough so that the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s will be relatively quite close to the true parameters.

The most obvious formula for determining $\hat{c}_{ijk}$ would be

$$(11.2) \qquad \hat{c}_{ijk} = \hat{\nu} T_{ijk} + \hat{a}_i + \hat{b}_i H_{ijk} \quad .$$

If the third approach was used to estimate the equating parameters, then the $\hat{a}_i$'s, the $\hat{b}_i$'s, and $\hat{\nu}$ of (11.2) will be respectively the estimates of the a_i's, the b_i's, and ν under the model (9.2) as calculated by the methods of Section 9. If the special form of the second approach was used to estimate the equating parameters, then the $\hat{\nu}$ of (11.2) will be the estimate of ν under the model (7.18), and the $\hat{a}_i$'s and $\hat{b}_i$'s of (11.2) will be respectively the estimates (7.5) and (7.6) multiplied by the estimate of ν° under (7.18). Appropriate modifications in (11.2) can be made if there is more than one T-variable or if the b_i's have been eliminated.

Now

$$(11.3) \qquad E(c_{ij} | T_{ijk}, H_{ijk}) = \nu T_{ijk} + a_i + b_i H_{ijk} \quad ,$$

where ν , a_i, and b_i are the parameters appearing in (9.2). Hence the expectation of

$$(11.4) \qquad c_{ijk} - \hat{c}_{ijk}$$

conditional upon T_{ijk} and H_{ijk} should for practical purposes be 0[since $\hat{a}_i$ and $\hat{b}_i$ as well as $\hat{\nu}$ in (11.2) ought to be virtually unbiased estimates of the parameters which they estimate]. The variance of (11.4) conditional upon T_{ijk} and H_{ijk} is approximately[16]

$$(11.5) \qquad \sigma^2_{(c-\hat{c})} = 1 + \frac{1}{N_{(i).}} + \frac{(H_{ijk} - \bar{H}_{(i).})^2}{S_{HH(i).}}$$

if the third approach was used; the corresponding formula for the special form of the second approach is treated in the Appendix[16]. The parentheses around the i's in (11.5) indicate that the associated quantities refer to the

data that was used to estimate the a_i's and b_i's, not to the data of the students for whom the $\hat{c}_{ijk}$'s are to be obtained.

Finally, we see that a 95% confidence interval for c_{ijk} (the equated college grade which is being predicted) is given approximately by

$$(11.6) \qquad \hat{c}_{ijk} \pm 1.96\, \sigma_{(c-\hat{c})} \quad ,$$

where $\hat{c}_{ijk}$ is specified by (11.2) and $\sigma_{(c-\hat{c})}$ is obtained from (11.5) [or from (A16.5]. Note from (11.5) that, in general, the interval (11.6) will be wider the smaller $N_{(i).}$ is.

For students from schools with very small $N_{(i).}$'s, (11.5) may become rather large, and we might want to consider whether it would be better in such cases to base $\hat{c}_{ijk}$ on T_{ijk} alone rather than on both T_{ijk} and H_{ijk}. In fact, $\sigma^2_{(c-\hat{c})}$ (11.5) is taken to be infinite if $N_{(i).}$ is 1 or 0.

We investigate the matter of basing $\hat{c}_{ijk}$ on T_{ijk} alone. Note first that the conditional distribution of c_{ijk} given T_{ijk} has mean of the form

$$(11.7) \qquad E(c_{ijk}|T_{ijk}) = \mu + \nu' T_{ijk}$$

[or, equivalently,

$$(11.8) \qquad E(c_{ijk}|T_{ijk}) = \mu_c + \rho_{CT} \frac{\sigma_c}{\sigma_T} (T_{ijk} - \mu_T) \quad ,$$

where the notation of (11.8) is explained in Note 15 of the Appendix], and variance

$$(11.9) \qquad \begin{aligned} \mathrm{var}(c_{ijk}|T_{ijk}) = \mathrm{var}(c|T) &= \sigma_c^2(1-\rho_{CT}^2) \\ &= \frac{(1-\rho_{CT}^2)(1-\rho_{TH}^2)}{(1-\rho_{CT}^2-\rho_{CH}^2-\rho_{TH}^2+2\rho_{CT}\rho_{CH}\rho_{TH})} \quad . \end{aligned}$$

[In obtaining the second line of (11.9), we assume that the c_{ijk}'s of the model (11.7-11.8) are based on the same β_j's as in (9.2), and then we utilize

the fact that (A15.5) is equal to 1, solve for σ_c^2 , and substitute into the expression $\sigma_c^2(1-\rho_{CT}^2)$ to get the final expression in (11.9). In this way, we have a formula which may be more suitable for any comparison involving (11.5)].

We suppose that we can obtain estimates of μ and ν' in (11.7), and of var(c|T)(11.9). [Such estimates under the model (11.7) would have to be calculated from data for which there was no selection on the basis of H; in fact, the data of students from the high schools having very small $N_{(i).}$'s and therefore presumably uninterpretable H-scores might be used, since such uninterpretable H-scores should not have been utilized for selection. Before starting any calculating for estimates under the model (11.7), values of the equating parameters for the college grades could be taken from the results of the main estimation procedure and used to transform the C_{ijk}'s in the data to c_{ijk}'s.] Once the estimates $\hat{\mu}$ and $\hat{\nu}'$ of μ and ν' have been found, we can write

$$\hat{c}_{ijk} = \hat{\mu} + \hat{\nu}'T_{ijk} \qquad (11.10)$$

as the formula for predicting c_{ijk} on the basis of T_{ijk} alone.

If $\hat{c}_{ijk}$ is given by (11.10), we see that the expectation of $(c_{ijk}-\hat{c}_{ijk})$ conditional upon T_{ijk} (but not the expectation conditional upon T_{ijk} and H_{ijk}) should for practical purposes be 0. Then, if we assume that the sample that was used for estimating μ and ν' was large enough so that the differences between $\hat{\mu}$ and μ and between $\hat{\nu}'$ and ν' are of relatively minor magnitude, we find that the variance of $(c_{ijk}-\hat{c}_{ijk})$ conditional upon T_{ijk} is approximately

$$\sigma^2_{(c-\hat{c})} = \mathrm{var}(c|T) \qquad , \qquad (11.11)$$

where the right-hand side of(11.11) may also be written in the alternate forms given by (11.9). Finally, we can use (11.10) and (11.11) to obtain a confidence interval like (11.6).

If we want to choose between the two prediction formulas (11.2) and (11.10), we might compare (11.5) [or (A16.5)] with (11.11) and decide according to which is smaller. However, we probably should use the same $\hat{c}_{ijk}$ formula for all the students of a given high school, rather than (e.g.) using (11.2) for those students whose H_{ijk}'s are sufficiently moderate that (11.11) exceeds (11.5) while using (11.10) for students whose H_{ijk}'s are extreme enough that (11.5) exceeds (11.11). If we used the latter procedure, then we might create a situation where two students from the same high school would have identical T-scores but the student with the higher H-score would have the lower $\hat{c}_{ijk}$; it would be hopeless to try to explain such an outcome to a college administrator.

Instead of making the comparison between (11.5) and (11.11) for each student individually, we could, as one possibility, compute the average of (11.5) across all $N_{i.}$ students in a given high school i, and then choose formula (11.2) or (11.10) for all students in high school i according as this average value of (11.5) is (respectively) smaller or larger than (11.11). A different possibility would be to use an approximation to this average rather than computing its exact value: the ratio of the numerator of the third term on the right-hand side of (11.5) to its denominator should, on the average, be roughly $[1+(1/N_{(i).})]$ to $[N_{(i).}-1]$, so that

$$(11.12) \qquad \sigma^2_{(c-\hat{c})} = 1 + \frac{1}{N_{(i).}} + \frac{1+(1/N_{(i).})}{N_{(i).}-1}$$

$$= 1 + \frac{2}{N_{(i).}-1}$$

represents an approximation to the average value of (11.5) for high school i. Now note that if we choose (11.2) or (11.10) according as (11.12) is (respectively) smaller or larger than (11.11), then our choice will be deter-

mined strictly by $N_{(i).}$, in such a way that we use (11.10) for all high schools with $N_{(i).}$'s below a certain number and we use (11.2) for all high schools with larger $N_{(i).}$'s.

The question might next be raised as to whether we could get a more sophisticated $\hat{c}_{ijk}$ by some kind of joint utilization of (11.2) and (11.10), rather than by using one or the other alone. For example, we might consider a prediction formula of the form

$$(11.13) \qquad \hat{c}_{ijk} = p(\hat{\nu}T_{ijk}+\hat{a}_i+\hat{b}_i H_{ijk})+(1-p)(\hat{\mu}+\hat{\nu}'T_{ijk}) \quad ,$$

which is a linear combination of (11.2) and (11.10). The coefficient p in (11.13) is to be between 0 and 1. We might determine p strictly on the basis of $N_{(i).}$, in which case p should increase as $N_{(i).}$ increases. If a $\sigma^2_{(c-\hat{c})}$ can be determined for (11.13), then this $\sigma^2_{(c-\hat{c})}$ should be a quadratic function in p, whose minimum with respect to p would be easily obtainable. However, some complications seem to arise when certain expectations pertaining to c_{ijk} and $\hat{c}_{ijk}$ are taken. In particular, we are faced in the beginning with the fact that, if $\hat{c}_{ijk}$ is given by (11.13), then $E[(c_{ijk}-\hat{c}_{ijk})|T_{ijk},H_{ijk}]$ is a linear function of T_{ijk} and H_{ijk} rather than being equal (approximately) to 0, and incidentally $E[(c_{ijk}-\hat{c}_{ijk})|\ T_{ijk}]$ is not 0 either unless we use for $E(H_{ijk}|\ T_{ijk})$ the same formula which holds in the unselected population (which probably would not be reasonable). If this paradox and other difficulties can be resolved, then the prediction formula (11.13) would appear to be a logical type of formula; but here in this report we will not attempt to explore its possibilities any further. For practical purposes, it would appear that (11.2) should be adequate for all students except a few coming from high schools with very small $N_{(i).}$'s, and for these latter students (11.10) can be used.

MATHEMATICAL APPENDIX

Note 1

The logarithm of the product over i, j, k of the expressions (5.4) is

$$\text{(A1.1)} \qquad L = \text{constant} + \sum_j N_{.j} \log|\beta_j| - \tfrac{1}{2} \sum_i \sum_j \sum_k (\alpha_j + \beta_j C_{ijk} - \nu T_{ijk})^2 \quad .$$

Differentiating L(A1.1) with respect to α_j, β_j, and ν, and setting the derivatives equal to 0, we obtain respectively

$$\text{(A1.2)} \qquad -N_{.j}\alpha_j - \beta_j C_{.j} + \nu T_{.j} = 0 \quad ,$$

$$\text{(A1.3)} \qquad (N_{.j}/\beta_j) - \alpha_j C_{.j} - \beta_j \sum_i \sum_k C_{ijk}^2 + \nu \sum_i \sum_k C_{ijk} T_{ijk} = 0 \quad ,$$

and

$$\text{(A1.4)} \qquad \sum_j \alpha_j T_{.j} + \sum_j \beta_j \sum_i \sum_k C_{ijk} T_{ijk} - \nu \sum_i \sum_j \sum_k T_{ijk}^2 = 0 \quad .$$

Upon solving (A1.2) for α_j, we get (5.8) (after putting on the hats). After substituting this solution for α_j into (A1.3), we end up with the quadratic equation

$$\text{(A1.5)} \qquad S_{CC.j}\beta_j^2 - \nu S_{CT.j}\beta_j - N_{.j} = 0$$

in β_j, which of course has two solutions. Since we can assume that $S_{CC.j}$ and $N_{.j}$ are both always > 0, one solution for β_j will always be positive and the other negative. One solution of (A1.5) is given by (5.7), and the second solution is the same thing except with the first plus in (5.7) replaced by a minus. The root (5.7) may be either positive or negative (according as $S_{CT.j}$ is positive or negative), but in either case it can be shown that (5.7) results in a larger value of L(A1.1) than does the other root, no matter what the value of $\nu (\nu > 0)$. For all practical purposes we can assume that (5.9) holds for each

j, so that all $\hat{\beta}_j$'s (5.7) will be > 0.

Substitution of the solutions for α_j and β_j [see (5.8) and (5.7)] into (A1.4) gives

$$\tfrac{1}{2}\nu \sum_j \frac{S^2_{CT.j}}{S_{CC.j}} + \tfrac{1}{2}\nu \sum_j \frac{S^2_{CT.j}}{S_{CC.j}} \left(1 + \frac{4N_{.j}}{S_{TT.j} r^2_j \nu^2}\right)^{\frac{1}{2}} - \nu \sum_j S_{TT.j} = 0 ,$$

which, after multiplication by $(-2/\nu)$ and application of the formula for r^2_j [see (2.4)], reduces to (5.6). Now there is one and only one value of ν^2 which satisfies (5.6). This is evident if we note that the left-hand side of (5.6) is a strictly increasing function of ν^2, becoming positive as $\nu^2 \to \infty$ (since r^2_j is ≤ 1, and can be assumed < 1), and becoming (infinitely) negative as $\nu^2 \to 0$. After finding the solution ν^2 of (5.6) (see Note 2), we take its positive square root to get $\hat{\nu}$. It can be shown that, mathematically speaking, no generality is gained if we permit negative values of ν.

Consider now the (2n+1)x (2n+1) matrix of second derivatives of L (A1.1) with respect to $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n, \nu$. We differentiate the left-hand sides of (A1.2), (A1.3), and (A1.4) to find the elements of this matrix. It can readily be seen that this matrix, after multiplication by -1, can be written as the sum of a matrix with the element $N_{.j}/\beta^2_j$ in the (n+j,n+j) diagonal position and zeroes everywhere else, plus a second matrix which is equal to the product of the (2n+1)x N.. matrix

(A1.6)

$$\left[\begin{array}{ccc|ccc|c|ccc}
1 & \cdots & 1 & 0 & \cdots & 0 & & 0 & \cdots & 0 \\
0 & \cdots & 0 & 1 & \cdots & 1 & & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & & 1 & \cdots & 1 \\
\hline
C_{111} & \cdots & C_{m1N_{m1}} & 0 & \cdots & 0 & & 0 & \cdots & 0 \\
0 & \cdots & 0 & C_{121} & \cdots & C_{m2N_{m2}} & & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & & C_{1n1} & \cdots & C_{mnN_{mn}} \\
\hline
-T_{111} & \cdots & -T_{m1N_{m1}} & -T_{121} & \cdots & -T_{m2N_{m2}} & & -T_{1n1} & \cdots & -T_{mnN_{mn}}
\end{array}\right]$$

by its transpose. Thus, since we can certainly assume that the last row of (A1.6) is not a linear combination of the first n rows, it follows that our matrix of second derivatives is negative definite for all values of the α_j's, the β_j's, and ν, including in particular those values which satisfy (A1.2-A1.4). This is sufficient to establish that any solution of the system (A1.2-A1.4) constitutes a relative maximum of the function L (A1.1) (see, e.g., Apostol [2, pp.151-152]). Furthermore, it can be shown that our solution based on (5.6-5.8) will constitute an absolute maximum (the second to the last paragraph of Note 3 below is partially relevant here). We have previously established the existence of this solution, and we have established that, if (5.9) holds for all j, then it is a unique solution of (A1.2-A1.4) under the restriction that ν and all β_j's are > 0.

Note 2

The Newton-Raphson method (see, e.g., [7, p.192 ff.]) is an iterative method of solving an equation of the form $f(z) = 0$: it uses the iteration formula

$$z_{new} = z_{old} - \frac{f(z_{old})}{f'(z_{old})} \tag{A2.1}$$

to find a new (and usually better) approximation to the root of $f(z) = 0$ at each step. Applying this formula (A2.1) to the equation (5.6), with $z = \nu^2$, we obtain the iteration formula

$$\nu^2_{new} = \nu^2_{old} - \frac{\sum_j S_{TT.j}\left[2-r_j^2-r_j^2\left(1+\frac{4N_{.j}}{S_{TT.j}r_j^2\nu^2_{old}}\right)^{\frac{1}{2}}\right]}{\frac{2}{(\nu^2_{old})^2}\sum_j \frac{N_{.j}}{\left(1+\frac{4N_{.j}}{S_{TT.j}r_j^2\nu^2_{old}}\right)^{\frac{1}{2}}}} \tag{A2.2}$$

for finding the root of (5.6).

There are certain sufficient conditions for the Newton-Raphson procedure (A2.1) to converge to the root. In particular, convergence is assured if $f'(z) > 0$ and $f''(z) < 0$ for all $z \geq$ the value used in the first iteration and $\leq$ the root (it being assumed that a starting value can be found which is to the left of the root). Now the denominator of the last term of (A2.2) is the first derivative of the left-hand side of (5.6) with respect to ν^2, and it is clearly positive for all $\nu^2 > 0$; the second derivative with respect to ν^2 is easily shown to be negative for all $\nu^2 > 0$. Hence convergence of the procedure (A2.2) is assured once we find a (positive) value of ν^2 which is smaller than the root.

Note 3

We treat the special case where there are just two T-variables; this will be sufficient to put across the general idea for larger numbers of T-variables. Let the two T-variables be denoted by T_{ijk} and T°_{ijk}. We assume a model in which the conditional distribution of c_{ijk} given T_{ijk} and T°_{ijk} is of the form

$$(2\pi)^{-\frac{1}{2}} \; e^{-\frac{1}{2}(c_{ijk}-\mu-\nu T_{ijk}-\nu^{\circ}T^{\circ}_{ijk})^2} \quad ,$$

so that [if we proceed analogously to (5.2-5.5)] the conditional distribution of C_{ijk} given T_{ijk} and T°_{ijk} is of the form

$$(A3.1) \qquad (2\pi)^{-\frac{1}{2}} |\beta_j| \; e^{-\frac{1}{2}(\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-\nu^{\circ}T^{\circ}_{ijk})^2} \quad .$$

The logarithm of the product over i,j,k of the expressions (A3.1) is

$$(A3.2) \qquad L = \text{const} + \sum_j N_{\cdot j} \log|\beta_j| - \tfrac{1}{2} \sum_i \sum_j \sum_k (\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-\nu^{\circ}T^{\circ}_{ijk})^2 \quad .$$

Differentiating L (A3.2) with respect to the α_j's, the β_j's, ν, and ν°,

and setting the derivatives equal to 0, we obtain the system

$$(A3.3) \qquad -N_{.j}\alpha_j - \beta_j C_{.j} + \nu T_{.j} + \nu^\circ T^\circ_{.j} = 0 \quad ,$$

$$(A3.4) \qquad (N_{.j}/\beta_j) - \alpha_j C_{.j} - \beta_j \sum_i \sum_k C^2_{ijk} + \nu \sum_i \sum_k C_{ijk} T_{ijk} + \nu^\circ \sum_i \sum_k C_{ijk} T^\circ_{ijk} = 0 \quad ,$$

$$(A3.5) \qquad \sum_j \alpha_j T_{.j} + \sum_j \beta_j \sum_i \sum_k C_{ijk} T_{ijk} - \nu \sum_i \sum_j \sum_k T^2_{ijk} - \nu^\circ \sum_i \sum_j \sum_k T_{ijk} T^\circ_{ijk} = 0 \quad ,$$

$$(A3.6) \qquad \sum_j \alpha_j T^\circ_{.j} + \sum_j \beta_j \sum_i \sum_k C_{ijk} T^\circ_{ijk} - \nu \sum_i \sum_j \sum_k T_{ijk} T^\circ_{ijk} - \nu^\circ \sum_i \sum_j \sum_k T^{\circ 2}_{ijk} = 0,$$

where $T^\circ_{.j}$ is defined analogously to $T_{.j}$. Let $\bar{T}^\circ_{.j}$, $S_{CT^\circ.j}$, $S_{T^\circ T^\circ.j}$, and $S_{TT^\circ.j}$ also be defined in analogy with previous notation [see (2.1, 2.4)]. Solving (A3.3) for α_j gives us

$$(A3.7) \qquad \hat{\alpha}_j = \hat{\nu}\,\bar{T}_{.j} + \hat{\nu}^\circ\,\bar{T}^\circ_{.j} - \hat{\beta}_j \bar{C}_{.j} \quad .$$

If we substitute this solution for α_j into (A3.4), we will obtain the quadratic equation

$$(A3.8) \qquad S_{CC.j}\,\beta_j^2 - (\nu\, S_{CT.j} + \nu^\circ S_{CT^\circ.j})\beta_j - N_{.j} = 0$$

in β_j. It can be shown that, whatever ν and ν° are, the root

$$(A3.9) \qquad \hat{\beta}_j = \frac{(\hat{\nu} S_{CT.j} + \hat{\nu}^\circ S_{CT^\circ.j})}{2S_{CC.j}} \left[1 + \left(1 + \frac{4N_{.j} S_{CC.j}}{(\hat{\nu} S_{CT.j} + \hat{\nu}^\circ S_{CT^\circ.j})^2}\right)^{\frac{1}{2}}\right]$$

of (A3.8) will result in a larger value of L (A3.2) than the other root, so that we estimate β_j by (A3.9). Finally, if we plug (A3.7) and then (A3.9) into (A3.5) and (A3.6), we will end up with a system of two equations in the two unknowns ν and ν°:

$$(A3.10)\qquad \nu\left(2\sum_j S_{TT.j} - \sum_j \frac{S^2_{CT.j}}{S_{CC.j}}\right) + \nu^\circ\left(2\sum_j S_{TT^\circ.j} - \sum_j \frac{S_{CT.j}S_{CT^\circ.j}}{S_{CC.j}}\right)$$

$$- \sum_j (\nu S_{CT.j} + \nu^\circ S_{CT^\circ.j}) \frac{S_{CT.j}}{S_{CC.j}} \left[1 + \frac{4N_{.j}S_{CC.j}}{(\nu S_{CT.j} + \nu^\circ S_{CT^\circ.j})^2}\right]^{\frac{1}{2}} = 0$$

and

$$(A3.11)\qquad \nu\left(2\sum_j S_{TT^\circ.j} - \sum_j \frac{S_{CT.j}S_{CT^\circ.j}}{S_{CC.j}}\right) + \nu^\circ\left(2\sum_j S_{T^\circ T^\circ.j} - \sum_j \frac{S^2_{CT^\circ.j}}{S_{CC.j}}\right)$$

$$-\sum_j (\nu S_{CT.j} + \nu^\circ S_{CT^\circ.j}) \frac{S_{CT^\circ.j}}{S_{CC.j}} \left[1 + \frac{4N_{.j}S_{CC.j}}{(\nu S_{CT.j} + \nu^\circ S_{CT^\circ.j})^2}\right]^{\frac{1}{2}} = 0 .$$

Once (A3.10-A3.11) are solved for ν and ν°, we can substitute the solutions $\hat{\nu}$ and $\hat{\nu}^\circ$ into (A3.9) and (A3.7) to obtain the $\hat{\beta}_j$'s and $\hat{\alpha}_j$'s. We will not attempt to explore the details of solving (A3.10-A3.11) for ν and ν°, but this problem can be attacked through methods of numerical analysis, such as the generalized Newton-Raphson method or the method of steepest descent.

In practical applications, all $\hat{\beta}_j$'s (A3.9) and all $S_{CT.j}$'s and $S_{CT^\circ.j}$'s will presumably be > 0. Furthermore, ν and ν° should be > 0. It may be helpful to realize that, if all $S_{CT.j}$'s and $S_{CT^\circ.j}$'s are > 0, then there can be no more than one solution of (A3.10-A3.11) such that ν and ν° are both > 0. This is an immediate consequence of (A3.9) and of a proposition which we now present.

We shall prove that there can be no more than one solution of (A3.3-A3.6) which lies in the set

$$(A3.12)\qquad \beta_j > 0 \text{ for all } j,\ -\infty < \alpha_j < \infty \text{ for all } j,\ -\infty < \nu,\ \nu^\circ < \infty \quad .$$

This will follow from the continuity of L (A3.2) and its derivatives in the set (A3.12), and from the fact that the matrix of second derivatives of L is negative definite throughout (A3.12). For, let $\underline{\chi}$ ([2n+2]x1) denote a vector consisting of $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n, \nu, \nu^\circ$. Suppose that $\underline{\chi}_1$ ([2n+2]x1) and $\underline{\chi}_2$([2n+2]x1) are two points in (A3.12) both satisfying (A3.3-A3.6); we show that this leads to a contradiction. Define $\underline{\delta}$ ([2n+2]x1) = $\underline{\chi}_2 - \underline{\chi}_1$. Define $\underline{L}_1(\underline{\chi})$ ([2n+2]x1) to be the vector of first derivatives of $L(\underline{\chi})$ (A3.2)[as given by the left-hand sides of (A3.3-A3.6)], and define $L_2(\underline{\chi})$ ([2n+2]x[2n+2]) to be the matrix of second derivatives of $L(\underline{\chi})$. $L_2(\underline{\chi})$ can be shown to be negative definite by using the same type of argument that was used in Note 1. Now define a function $g(\lambda) = L(\underline{\chi}_1 + \lambda \underline{\delta})$, where λ is a scalar. Since L and its derivatives are continuous in $\underline{\chi}$ throughout (A3.12), $g(\lambda)$ and its derivatives will be continuous for $0 \leq \lambda \leq 1$. We find $g'(\lambda) = \underline{\delta}'\underline{L}_1(\underline{\chi}_1 + \lambda \underline{\delta})$, so that $g'(0) = \underline{\delta}'\underline{L}_1(\underline{\chi}_1) = 0$ and $g'(1) = \underline{\delta}'\underline{L}_1(\underline{\chi}_2) = 0$. Also $g''(\lambda) = \underline{\delta}'L_2(\underline{\chi}_1 + \lambda \underline{\delta})\underline{\delta}$, so that $g''(\lambda) < 0$ for all λ in the interval $0 \leq \lambda \leq 1$ since L_2 is negative definite. But, by the law of the mean, $g'(0) = g'(1) = 0$ implies that $g''(\lambda)$ must be 0 for some λ between 0 and 1. Hence the contradiction.

Solving the system (A3.10-A3.11) is distinctly more complicated than solving the comparatively simple equation (5.6) via the iteration procedure (A2.2). Furthermore, the complications increase as the number of T-variables becomes larger. For the sake of completeness, we mention an alternative method of solving the system (A3.3-A3.6). After obtaining (A3.7), we substitute this solution for α_j into (A3.5-A3.6) and get a linear system in ν and ν°,

$$\text{(A3.13)} \qquad \nu \sum_j S_{TT.j} + \nu^\circ \sum_j S_{TT^\circ.j} = \sum_j \beta_j S_{CT.j}$$

$$(A3.14) \qquad \nu \sum_j S_{TT^\circ.j} + \nu^\circ \sum_j S_{T^\circ T^\circ.j} = \sum_j \beta_j S_{CT^\circ.j} \quad ,$$

which may be solved for ν and ν°. The resulting formulas for ν and ν°, which will be linear combinations of the β_j's, may be substituted [along with (A3.7)] into (A3.4) to obtain an equation system in the β_j's, which will be linear except for the $(N_{.j}/\beta_j)$ terms. We will not dwell here on the solution of this system in the β_j's, except to say that a similar system will arise later in connection with our third approach (Section 9 of the paper) and will be considered in detail.

Note 4

Actually, instead of working with the model (5.14) which has only one T-variable, we use a model with two T-variables,

$$(A4.1) \qquad E(c_{ijk}) = a_i + b_i H_{ijk} + bT_{ijk} + b^\circ T^\circ_{ijk} \quad ,$$

in order to provide a more general development. We shall use notation for the c_{ijk}'s (and T°_{ijk}'s) which is in analogy with (2.1) and (2.4); remember, though, that in an application of the type being considered in this report, it is really the c^*_{ijk}'s (5.13) rather than the c_{ijk}'s which will have to be entered into the calculations.

Our object is to obtain the least-squares estimates (which are the same as the maximum-likelihood estimates in this case, if the distribution is normal) of the a_i's, the b_i's, b, and b°. First we re-write (A4.1) in the form

$$(A4.2) \qquad E(c_{ijk}) = A_i + b_i(H_{ijk}-\bar{H}_{i.}) + b(T_{ijk}-\bar{T}_{i.}) + b^\circ(T^\circ_{ijk}-\bar{T}^\circ_{i.}) \quad ,$$

where

(A4.3) $$A_i = a_i + b_i\bar{H}_{i.} + b\,\bar{T}_{i.} + b^\circ\bar{T}^\circ_{i.}$$.

Then the normal equations are easily found to be

(A4.4)
$$\left[\begin{array}{ccc|ccc|cc}
N_{1.} & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
0 & \cdots & N_{m.} & 0 & \cdots & 0 & 0 & 0 \\
\hline
0 & \cdots & 0 & S_{HH1.} & \cdots & 0 & S_{TH1.} & S_{T^\circ H1.} \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & S_{HHm.} & S_{THm.} & S_{T^\circ Hm.} \\
\hline
0 & \cdots & 0 & S_{TH1.} & \cdots & S_{THm.} & \Sigma_i S_{TTi.} & \Sigma_i S_{TT^\circ i.} \\
0 & \cdots & 0 & S_{T^\circ H1.} & \cdots & S_{T^\circ Hm.} & \Sigma_i S_{TT^\circ i.} & \Sigma_i S_{T^\circ T^\circ i.}
\end{array}\right]
\left[\begin{array}{c} A_1 \\ \vdots \\ A_m \\ \hline b_1 \\ \vdots \\ b_m \\ \hline b \\ b^\circ \end{array}\right]
=
\left[\begin{array}{c} c_{1.} \\ \vdots \\ c_{m.} \\ \hline S_{cH1.} \\ \vdots \\ S_{cHm.} \\ \hline \Sigma_i S_{cTi.} \\ \Sigma_i S_{cT^\circ i.} \end{array}\right].$$

Thus $\hat{A}_{i.} = \bar{c}_{i.}$, so that

(A4.5) $$\hat{a}_i = \bar{c}_{i.} - \hat{b}_i\bar{H}_{i.} - \hat{b}\bar{T}_{i.} - \hat{b}^\circ\bar{T}^\circ_{i.}$$

in view of (A4.3). To obtain $(\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_m, \hat{b}, \hat{b}^\circ)$, we need to invert the (m+2)x(m+2) sub-matrix appearing in the lower right-hand corner of the matrix in (A4.4), and then multiply the resulting inverse by the (m+2)x1 sub-vector which comprises the bottom (m+2) positions of the vector on the right-hand side of (A4.4). This (m+2)x(m+2) matrix which has to be inverted we may denote by V. We then define certain additional matrices by the identifications

(A4.6) $$V = \left[\begin{array}{c|c} V_{HH}(m\text{x}m) & V_{TH}(m\text{x}2) \\ \hline V'_{TH}(2\text{x}m) & V_{TT}(2\text{x}2) \end{array}\right]$$

and

(A4.7) $$V^{-1} = \left[\begin{array}{c|c} V^{HH}(m\text{x}m) & V^{TH}(m\text{x}2) \\ \hline V^{TH'}(2\text{x}m) & V^{TT}(2\text{x}2) \end{array}\right]$$.

Now

$$(A4.8a) \qquad V^{TT} = (V_{TT} - V'_{TH}\, V_{HH}^{-1}\, V_{TH})^{-1} \quad ,$$

$$(A4.8b) \qquad V^{TH} = -V_{HH}^{-1} V_{TH} V^{TT} \quad ,$$

and

$$(A4.9c) \qquad V^{HH} = (I - V^{TH} V'_{TH}) V_{HH}^{-1} \quad ,$$

where I denotes the $(m \times m)$ identity matrix. Formulas (A4.8) are standard formulas which are used in connection with the inversion of partitioned matrices; their correctness can easily be verified directly, however. Observe that these formulas (A4.8) do not require the inversion of any matrices of higher order than 2 (except for the matrix V_{HH}, but V_{HH} is diagonal). Thus V^{-1} is found rather easily by performing the three calculations (A4.8) and then plugging the resulting matrices into (A4.7). We can then obtain all the desired estimates.

In case we are willing to simplify our model by eliminating the b'_i's (i.e., setting them equal to 1) in (5.14), then (5.14) is altered by replacing b_i with b'. With two T-variables [as in (A4.1)], the model takes the form

$$(A4.9) \qquad E(c_{ijk}) = a_i + b'H_{ijk} + bT_{ijk} + b^\circ T^\circ_{ijk} \quad .$$

The parameters are then estimated by the formulas

$$(A4.10) \qquad \hat{a}_i = \bar{c}_{i.} - \hat{b}'\bar{H}_{i.} - \hat{b}\bar{T}_{i.} - \hat{b}^\circ \bar{T}^\circ_{i.}$$

and

$$(A4.11) \qquad \begin{bmatrix} \hat{b}' \\ \hat{b} \\ \hat{b}^\circ \end{bmatrix} = \begin{bmatrix} \sum_i S_{HHi.} & \sum_i S_{THi.} & \sum_i S_{T^\circ Hi.} \\ \sum_i S_{THi.} & \sum_i S_{TTi.} & \sum_i S_{TT^\circ i.} \\ \sum_i S_{T^\circ Hi.} & \sum_i S_{TT^\circ i.} & \sum_i S_{T^\circ T^\circ i.} \end{bmatrix}^{-1} \begin{bmatrix} \sum_i S_{cHi.} \\ \sum_i S_{cTi.} \\ \sum_i S_{cT^\circ i.} \end{bmatrix} \quad .$$

Note 5

We consider explicitly just the case where there are exactly two T-variables, T_{ijk} and T°_{ijk}. The extension of the theory to the case of more than two T-variables will be obvious, however.

If the joint distribution of T_{ijk}, T°_{ijk}, and h_{ijk} is tri-variate normal in the unselected population, then the conditional joint distribution of T_{ijk} and T°_{ijk} given h_{ijk}, for either the unselected or the selected population, will be (see, e.g., [1], p.29, Theorem 2.5.1) a bivariate normal distribution of the form

$$(A5.1) \qquad (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\begin{pmatrix} T_{ijk}-\mu-b(a_i+b_iH_{ijk}) \\ T^{\circ}_{ijk}-\mu^{\circ}-b^{\circ}(a_i+b_iH_{ijk}) \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} T_{ijk}-\mu-b(a_i+b_iH_{ijk}) \\ T^{\circ}_{ijk}-\mu^{\circ}-b^{\circ}(a_i+b_iH_{ijk}) \end{pmatrix}}$$

after we have substituted for h_{ijk} in accordance with (2.3). Note that, in (A5.1), the (a_i, b_i)'s are not unique, in the sense that, if all a_i's and b_i's are multiplied by the same constant, this constant can be absorbed in b and b°, and if all a_i's are increased or decreased by the same constant, this adjustment can be absorbed in μ and μ°. We will show that maximum-likelihood estimates of these equating parameters a_i and b_i are given by

$$(A5.2) \qquad \hat{a}_i = N_{\cdot\cdot} \frac{(B,\ B^{\circ})\begin{pmatrix} \bar{T}_{i\cdot} - \bar{T}_{\cdot\cdot} \\ \bar{T}^{\circ}_{i\cdot} - \bar{T}^{\circ}_{\cdot\cdot} \end{pmatrix}}{(B,\ B^{\circ})\ (W_1 - W)\begin{pmatrix} B \\ B^{\circ} \end{pmatrix}} - \hat{b}_i \bar{H}_{i\cdot}$$

and

$$(A5.3) \qquad \hat{b}_i = \frac{N_{\cdot\cdot}}{S_{HHi\cdot}} \frac{(B,B^{\circ})\begin{pmatrix} S_{THi\cdot} \\ S_{T^{\circ}Hi\cdot} \end{pmatrix}}{(B,B^{\circ})(W_1 - W)\begin{pmatrix} B \\ B^{\circ} \end{pmatrix}} ,$$

where

(A5.4)

$$W(2x2)=\begin{bmatrix} \sum_i \frac{S^2_{THi.}}{S_{HHi.}} + \sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})^2 & \sum_i \frac{S_{THi.}S_{T^\circ Hi.}}{S_{HHi.}} + \sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..}) \\ \sum_i \frac{S_{THi.}S_{T^\circ Hi.}}{S_{HHi.}} + \sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..}) & \sum_i \frac{S^2_{T^\circ Hi.}}{S_{HHi.}} + \sum_i N_{i.}(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..})^2 \end{bmatrix}$$

and

$$(A5.5) \qquad W_1(2x2) = \begin{bmatrix} S_{TT} & S_{TT^\circ} \\ S_{TT^\circ} & S_{T^\circ T^\circ} \end{bmatrix}$$

(S_{TT° and $S_{T^\circ T^\circ}$ being defined analogously to S_{TT}), so that

$$(A5.6) \quad W_1-W = \begin{bmatrix} \sum_i \left(S_{TTi.} - \frac{S^2_{THi.}}{S_{HHi.}}\right) & \sum_i \left(S_{TT^\circ i.} - \frac{S_{THi.}S_{T^\circ Hi.}}{S_{HHi.}}\right) \\ \sum_i \left(S_{TT^\circ i.} - \frac{S_{THi.}S_{T^\circ Hi.}}{S_{HHi.}}\right) & \sum_i \left(S_{T^\circ T^\circ i.} - \frac{S^2_{T^\circ Hi.}}{S_{HHi.}}\right) \end{bmatrix},$$

and where $(B, B^\circ)'$ is a characteristic vector of the matrix $(W_1-W)^{-1}W$ corresponding to the largest characteristic root of $(W_1-W)^{-1}W$. In other words, B and B° satisfy the equation

$$(A5.7) \qquad [(W_1-W)^{-1}W-\lambda I]\begin{pmatrix} B \\ B^\circ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where λ is the maximum root of $(W_1-W)^{-1}W$.

We now show how these estimators were derived. The logarithm of the product over (i,j,k) of the expressions (A5.1) is

$$(A5.8) \quad L = \text{const} + \tfrac{1}{2}N..\log|\Sigma^{-1}| - \tfrac{1}{2}\sum_i\sum_j\sum_k (T_{ijk}-\mu, T^\circ_{ijk}-\mu^\circ)\Sigma^{-1}\begin{pmatrix} T_{ijk}-\mu \\ T^\circ_{ijk}-\mu^\circ \end{pmatrix}$$

(formula continued on following page)

(A5.8) (cont.)

$$-\tfrac{1}{2}(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}\sum_i\sum_j\sum_k(a_i+b_iH_{ijk})^2+\sum_i\sum_j\sum_k(a_i+b_iH_{ijk})(b,b^\circ)\Sigma^{-1}\begin{pmatrix}T_{ijk}-\mu\\ T^\circ_{ijk}-\mu^\circ\end{pmatrix} .$$

Now if we differentiate L(A5.8) with respect to the elements of Σ (2x2) and then set the resulting derivatives equal to 0, we obtain the equation

$$(A5.9)\qquad N..\Sigma = \sum_i\sum_j\sum_k\begin{pmatrix}T_{ijk}-\mu-b(a_i+b_iH_{ijk})\\ T^\circ_{ijk}-\mu^\circ-b^\circ(a_i+b_iH_{ijk})\end{pmatrix}\begin{pmatrix}T_{ijk}-\mu-b(a_i+b_iH_{ijk})\\ T^\circ_{ijk}-\mu^\circ-b^\circ(a_i+b_iH_{ijk})\end{pmatrix}' ,$$

which is the condition for L to be maximized with respect to Σ (see, e.g., [1], pp.46-47, Lemma 3.2.2 for details). The other first derivatives of L(A5.8) are given by

$$(A5.10)\qquad \frac{\partial L}{\partial a_i} = -(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}(N_{i.}a_i+b_iH_{i.})+(b,b^\circ)\Sigma^{-1}\begin{pmatrix}T_{i.}-\mu\\ T^\circ_{i.}-\mu^\circ\end{pmatrix} ,$$

$$(A5.11)\qquad \frac{\partial L}{\partial b_i} = -(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}(a_iH_{i.}+b_i\sum_{jk}\sum H^2_{ijk})+(b,b^\circ)\Sigma^{-1}\begin{pmatrix}\sum_{jk}\sum T_{ijk}H_{ijk}-\mu H_{i.}\\ \sum_{jk}\sum T^\circ_{ijk}H_{ijk}-\mu^\circ H_{i.}\end{pmatrix},$$

$$(A5.12)\qquad \begin{pmatrix}\frac{\partial L}{\partial \mu}\\ \frac{\partial L}{\partial \mu^\circ}\end{pmatrix} = \Sigma^{-1}\begin{pmatrix}T..-N..\mu\\ T^\circ..-N..\mu^\circ\end{pmatrix}-\sum_i(N_{i.}a_i+b_iH_{i.})\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix} ,$$

and

$$(A5.13)\qquad \begin{pmatrix}\frac{\partial L}{\partial b}\\ \frac{\partial L}{\partial b^\circ}\end{pmatrix}=-\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}\sum_{ijk}\sum(a_i+b_iH_{ijk})^2+\sum_{ijk}\sum(a_i+b_iH_{ijk})\Sigma^{-1}\begin{pmatrix}T_{ijk}-\mu\\ T^\circ_{ijk}-\mu^\circ\end{pmatrix} .$$

Setting (A5.10) equal to 0 and solving for a_i, we obtain

$$(A5.14) \qquad a_i = \frac{(b,b^\circ)\,\Sigma^{-1}\begin{pmatrix}\bar{T}_{i.}-\mu\\ \bar{T}^\circ_{i.}-\mu^\circ\end{pmatrix}}{(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}} - b_i\bar{H}_{i.} \quad .$$

Next we set (A5.11) equal to 0, substitute from (A5.14) for a_i, and solve for b_i. We get

$$(A5.15) \qquad b_i = \frac{(b,b^\circ)\Sigma^{-1}\begin{pmatrix}S_{THi.}\\ S_{T^\circ Hi.}\end{pmatrix}}{S_{HHi.}\,(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}} \quad .$$

If we set both rows of (A5.12) equal to 0 and then plug in(A5.14), we see that

$$(A5.16) \qquad \mu = \bar{T}.. \quad , \qquad \mu^\circ = \bar{T}^\circ_{\cdot\cdot}$$

satisfies the resulting equation system. (The solution for μ and μ° is not unique, due to reasons mentioned previously.) Now we set both rows of (A5.13) equal to 0, pre-multiply by Σ, and substitute from (A5.14-A5.16) to obtain

$$(A5.17) \qquad -\frac{(b,b^\circ)\,\Sigma^{-1}W\Sigma^{-1}(b,b^\circ)'}{[(b,b^\circ)\Sigma^{-1}(b,b^\circ)']^2}\begin{pmatrix}b\\ b^\circ\end{pmatrix} + \frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'}W\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad ,$$

where W is given by (A5.4). If we substitute (A5.14-A5.16) into (A5.9), then (A5.9) reduces to

$$(A5.18) \qquad N..\Sigma = W_1 + \frac{(b,b^\circ)\Sigma^{-1}W\Sigma^{-1}(b,b^\circ)'}{[(b,b^\circ)\Sigma^{-1}(b,b^\circ)']^2}\begin{pmatrix}b\\ b^\circ\end{pmatrix}(b,b^\circ)$$

$$- \frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'}\left[W\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}(b,b^\circ) + \begin{pmatrix}b\\ b^\circ\end{pmatrix}(b,b^\circ)\;\Sigma^{-1}W\right] \quad ,$$

where W_1 is given by (A5.5). Now (A5.18) reduces further to

$$(A5.19) \qquad N..\Sigma = W_1 - \frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'} W\Sigma^{-1}\begin{pmatrix} b \\ b^\circ \end{pmatrix}(b,b^\circ)$$

upon substitution of the relation (A5.17). Thus we obtain

$$(A5.20) \qquad N..\begin{pmatrix} b \\ b^\circ \end{pmatrix} = (W_1 - W)\,\Sigma^{-1}\begin{pmatrix} b \\ b^\circ \end{pmatrix}$$

upon post-multiplying (A5.19) by $\Sigma^{-1}(b, b^\circ)'$, and

$$(A5.21) \qquad N..(b,b^\circ)\Sigma^{-1}\begin{pmatrix} b \\ b^\circ \end{pmatrix} = N..(B,B^\circ)\Sigma\begin{pmatrix} B \\ B^\circ \end{pmatrix} = (B,B^\circ)(W_1 - W)\begin{pmatrix} B \\ B^\circ \end{pmatrix}$$

after pre-multiplying (A5.20) by $(b,b^\circ)\Sigma^{-1}$ and then substituting

$$(A5.22) \qquad \begin{pmatrix} B \\ B^\circ \end{pmatrix} = \Sigma^{-1}\begin{pmatrix} b \\ b^\circ \end{pmatrix} \quad ,$$

where (A5.22) defines B and B°. After application of the relations (A5.20-A5.22), equation (A5.17) will finally reduce to

$$(A5.23) \qquad \left[W - \frac{(B,B^\circ)W(B,B^\circ)'}{(B,B^\circ)(W_1 - W)(B,B^\circ)'}(W_1 - W)\right]\begin{pmatrix} B \\ B^\circ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad .$$

Now the only way for $(B,B^\circ)'$ to satisfy (A5.23) is for it to be a characteristic vector of the matrix $(W_1 - W)^{-1}W$. In order to maximize L(A5.8), we must choose a characteristic vector corresponding to the largest characteristic root λ of $(W_1 - W)^{-1}W$, as indicated by (A5.7). Note finally that, upon substitution of (A5.16), (A5.21), and (A5.22) into (A5.14-A5.15), we obtain (A5.2-A5.3).

We might wish to consider what happens under a simplified model in which the b_i's in (A5.1) are all eliminated and set equal to 1. It turns out

that estimation under this model is not really too much easier than under the model (A5.1). We will show that the estimator for a_i is given by the formula

$$(A5.24) \qquad \hat{a}_i = \left(\sum_{i'} S_{HHi'.}\right) \frac{(B,B^\circ)\begin{pmatrix}\bar{T}_{i.}-\bar{T}_{..}\\ \bar{T}^\circ_{i.}-\bar{T}^\circ_{..}\end{pmatrix}}{(B,B^\circ)\begin{pmatrix}\sum_{i'} S_{THi'.}\\ \sum_{i'} S_{T^\circ Hi'.}\end{pmatrix}} - \bar{H}_{i.} \quad ,$$

where $(B,B^\circ)'$ is a characteristic vector of the matrix $W_3^{-1}W_4$ corresponding to the largest root of this matrix. Here we are defining

$$(A5.25) \qquad W_3(2x2) = \frac{1}{N..}\left[W_1 - W_2 - \frac{1}{\sum_i S_{HHi.}}\begin{pmatrix}\sum_i S_{THi.}\\ \sum_i S_{T^\circ Hi.}\end{pmatrix}\left(\sum_i S_{THi.}, \sum_i S_{T^\circ Hi.}\right)\right]$$

and

$$(A5.26) \qquad W_4(2x2) = W_2 + \frac{1}{\sum_i S_{HHi.}}\begin{pmatrix}\sum_i S_{THi.}\\ \sum_i S_{T^\circ Hi.}\end{pmatrix}\left(\sum_i S_{THi.}, \sum_i S_{T^\circ Hi.}\right) = W_1 - N..W_3 \quad ,$$

where W_1 is given by (A5.5) and

$$(A5.27) \qquad W_2(2x2) = \begin{bmatrix}\sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})^2 & \sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..})\\ \sum_i N_{i.}(\bar{T}_{i.}-\bar{T}_{..})(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..}) & \sum_i N_{i.}(\bar{T}^\circ_{i.}-\bar{T}^\circ_{..})^2\end{bmatrix} .$$

Thus B and B° satisfy the equation

$$(A5.28) \qquad \left[W_3^{-1}W_4 - \lambda' I\right]\begin{pmatrix}B\\ B^\circ\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad ,$$

where λ' is the maximum characteristic root of $W_3^{-1}W_4$.

The derivation of the estimators is somewhat similar to the derivation

for the previous model. Equations (A5.1), (A5.8-A5.10), (A5.12-A5.14), and (A5.16) go through just as before, except with all b_i's replaced by 1. Thus the solution for a_i [corresponding to (A5.14)] becomes

$$(A5.29) \qquad a_i = \frac{(b,b^\circ)\Sigma^{-1}\begin{pmatrix}\bar{T}_{i.}-\bar{T}_{..}\\ \bar{T}^\circ_{i.}-\bar{T}^\circ_{.}\end{pmatrix}}{(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}} - \bar{H}_{i.}$$

after substitution of (A5.16). After this point, the solution of the maximum-likelihood equations takes a somewhat different course. Instead of (A5.17) we obtain

$$(A5.30) \qquad - \left[\frac{(b,b^\circ)\Sigma^{-1}W_2\Sigma^{-1}(b,b^\circ)'}{[(b,b^\circ)\Sigma^{-1}(b,b^\circ)']^2} + \sum_i S_{HHi.}\right]\begin{pmatrix}b\\ b^\circ\end{pmatrix}$$

$$+ \frac{1}{(b,b^\circ)\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}} W_2\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix} + \begin{pmatrix}\sum_i S_{THi.}\\ \sum_i S_{T^\circ Hi.}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} ,$$

and instead of (A5.18) we have

$$(A5.31) \qquad N_{..}\Sigma = W_1 + \left[\frac{(b,b^\circ)\Sigma^{-1}W_2\Sigma^{-1}(b,b^\circ)'}{[(b,b^\circ)\Sigma^{-1}(b,b^\circ)']^2} + \sum_i S_{HHi.}\right]\begin{pmatrix}b\\ b^\circ\end{pmatrix}(b,b^\circ)$$

$$- \begin{pmatrix}b\\ b^\circ\end{pmatrix}\left[\frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'}(b,b^\circ)\Sigma^{-1}W_2 + \left(\sum_i S_{THi.}, \sum_i S_{T^\circ Hi.}\right)\right]$$

$$- \left[\frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'} W_2\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix} + \begin{pmatrix}\sum_i S_{THi.}\\ \sum_i S_{T^\circ Hi.}\end{pmatrix}\right](b,b^\circ) .$$

If the relation (A5.30) is applied to (A5.31), then (A5.31) reduces to

$$(A5.32) \qquad N_{..}\Sigma = W_1 - \frac{1}{(b,b^\circ)\Sigma^{-1}(b,b^\circ)'} W_2\Sigma^{-1}\begin{pmatrix}b\\ b^\circ\end{pmatrix}(b,b^\circ) - \begin{pmatrix}\sum_i S_{THi.}\\ \sum_i S_{T^\circ Hi.}\end{pmatrix}(b,b^\circ) .$$

If (A5.30) is pre-multiplied by (B,B°), where B and B° are again defined by (A5.22), we will arrive at the equation

$$(A5.33) \qquad (b,b^\circ)\Sigma^{-1}\begin{pmatrix} b \\ b^\circ \end{pmatrix} = \frac{1}{\sum_i S_{HHi.}} \left(\sum_i S_{THi.}, \sum_i S_{T^\circ Hi.}\right)\begin{pmatrix} B \\ B^\circ \end{pmatrix} .$$

Now if (A5.32) is post-multiplied by $(B,B^\circ)'$, then we will end up with the relation

$$(A5.34) \qquad \begin{pmatrix} b \\ b^\circ \end{pmatrix} = W_3 \begin{pmatrix} B \\ B^\circ \end{pmatrix}$$

after substituting (A5.33) and dividing by $N..$. Observe that from (A5.26), (A5.33), and (A5.34) it follows that

$$(A5.35) \qquad (B,B^\circ)W_4(B,B^\circ)' = (B,B^\circ)W_2(B,B^\circ)' + \sum_i S_{HHi.}\left[(B,B^\circ)W_3(B,B^\circ)'\right]^2 .$$

Thus (A5.30) becomes

$$(A5.36) \qquad \left[W_4 - \frac{(B,B^\circ)\, W_4(B,B^\circ)'}{(B,B^\circ)\, W_3(B,B^\circ)'}\, W_3 \right]\begin{pmatrix} B \\ B^\circ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

upon multiplying by $(b,b^\circ)\,\Sigma^{-1}\,(b,b^\circ)'$ and then applying (A5.33 -A5.35). From (A5.36) it follows that $(B,B^\circ)'$ must be a characteristic vector of $W_3^{-1}W_4$; in order to maximize the likelihood, we choose a characteristic vector corresponding to the largest root λ' , as indicated by (A5.28). This determines (B,B°) except for a multiplicative constant, but it is obviously not necessary to find this multiplicative constant for purposes of using formula (A5.24), which is obtained from (A5.29) by substituting (A5.33).

Note 6

The logarithm of the product over (i,j,k) of the expressions (7.11) is

$$(A6.1)\quad L = \text{const} - \tfrac{1}{2} N..\log(1-\rho^2) + N..\log\theta + \sum_i N_{i.} \log|b_i| - \frac{1}{2(1-\rho^2)} \sum_{ijk} \left[\theta^2 (T_{ijk}-\mu)^2 + (a_i+b_i H_{ijk})^2 - 2\rho\theta (T_{ijk}-\mu)(a_i+b_i H_{ijk}) \right] .$$

We differentiate L (A6.1) with respect to a_i, b_i, and μ, and obtain

$$(A6.2)\quad \frac{\partial L}{\partial a_i} = \frac{1}{1-\rho^2} \left[-(N_{i.} a_i + b_i H_{i.}) + \rho\theta (T_{i.} - N_{i.}\mu) \right] ,$$

$$(A6.3)\quad \frac{\partial L}{\partial b_i} = \frac{N_{i.}}{b_i} + \frac{1}{1-\rho^2} \left[-(a_i H_{i.} + b_i \sum_{jk} H_{ijk}^2) + \rho\theta (\sum_{jk} T_{ijk} H_{ijk} - \mu H_{i.}) \right] ,$$

and

$$(A6.4)\quad \frac{\partial L}{\partial \mu} = \frac{1}{1-\rho^2} \left[\theta^2 (T.. - N..\mu) - \rho\theta (\sum_i N_{i.} a_i + \sum_i b_i H_{i.}) \right] .$$

If we set (A6.2) equal to 0, solve for $(N_{i.} a_i + b_i H_{i.})$, sum over i, and substitute the result into (A6.4) after setting (A6.4) equal to 0, then we find

$$(A6.5)\quad \hat{\mu} = \bar{T}..$$

upon solving for μ. After setting (A6.2) equal to 0 and substituting (A6.5), we solve for a_i and end up with (7.12). Next we put (A6.3) equal to 0, substitute (7.12) and (A6.5), and obtain eventually the quadratic equation

$$(A6.6)\quad S_{HHi.} b_i^2 - \rho\theta S_{THi.} b_i - N_{i.}(1-\rho^2) = 0$$

in b_i. One solution of (A6.6) must be positive and the other negative. One solution is given by (7.13), and the other solution is the same thing except with the first plus in (7.13) replaced by a minus. In any case, the solution (7.13) will result in a larger value of L (A6.1) than the other solution. Note that (7.13) will be positive so long as $S_{THi.} > 0$ ($\hat{\rho}$ being assumed > 0).

The $(2m+1)x(2m+1)$ matrix of second derivatives of L (A6.1) with respect to $a_1, a_2, \ldots, a_m$, $b_1, b_2, \ldots, b_m$, μ is negative definite for all values of these parameters and all values of θ and ρ. This is easily established by an argument similar to the one used in Note 1. Thus we are able to conclude that, for any fixed values of θ and ρ, L (A6.1) takes on an absolute maximum when the a_i's, b_i's, and μ are as given by (7.12), (7.13), and (A6.5).

If we substitute (7.12), (7.13), and (A6.5) into the formula for L(A6.1), and also utilize (A6.6), then (with the constant terms omitted) L becomes the involved function of θ and ρ which is given by (7.14). Thus the maximum-likelihood estimates of θ and ρ are the values which maximize (7.14). After these values are found, they are plugged into (7.12) and (7.13) in order to get the maximum-likelihood estimates of the a_i's and b_i's.

Note 7

Actually, we will consider explicitly just the case where there are exactly two T-variables. From our development for two T-variables, the extension of the theory to the case of more than two T-variables will be obvious, and at the same time it will be apparent how to prove formula (7.16), which is for the case of just one T-variable. The analogue of (7.16) for two T-variables is given by (A7.13) below.

With two T-variables, we work with the tri-variate normal distribution

$$(A7.1)\quad (2\pi)^{-\frac{3}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(T_{ijk}-\mu,\ T^{\circ}_{ijk}-\mu^{\circ}, a_i+H_{ijk})\Sigma^{-1}\begin{pmatrix} T_{ijk}-\mu \\ T^{\circ}_{ijk}-\mu^{\circ} \\ a_i+H_{ijk}\end{pmatrix}\right]$$

in place of the distribution (7.15). We define certain matrices by the equations

$$(A7.2)\quad \Sigma(3\times3) = \begin{bmatrix} \Sigma_{TT}(2\times2) & \Sigma_{TH}(2\times1) \\ \Sigma'_{TH}(1\times2) & \Sigma_{HH}(1\times1) \end{bmatrix}, \quad \Sigma^{-1}(3\times3) = \begin{bmatrix} \Sigma^{TT}(2\times2) & \Sigma^{TH}(2\times1) \\ \Sigma^{TH'}(1\times2) & \Sigma^{HH}(1\times1) \end{bmatrix} .$$

The logarithm of the product over (i,j,k) of the expressions (A7.1) is then

$$(A7.3)\quad L = \text{const} - \tfrac{1}{2} N..\log|\Sigma| - \tfrac{1}{2} \sum_{ijk} \Big[(T_{ijk}-\mu, T^{\circ}_{ijk}-\mu^{\circ}) \Sigma^{TT} \begin{pmatrix} T_{ijk}-\mu \\ T^{\circ}_{ijk}-\mu^{\circ} \end{pmatrix} + (a_i+H_{ijk})^2 \Sigma^{HH} + 2(T_{ijk}-\mu, T^{\circ}_{ijk}-\mu^{\circ}) \Sigma^{TH}(a_i+H_{ijk}) \Big] .$$

After differentiating L (A7.3) with respect to the elements of Σ and then setting the resulting derivatives equal to 0, we obtain the equation (see e.g., [1], pp.46-47, Lemma 3.2.2 for details)

$$(A7.4)\quad N..\Sigma = \begin{bmatrix} \sum_{ijk}(T_{ijk}-\mu)^2 & \sum_{ijk}(T_{ijk}-\mu)(T^{\circ}_{ijk}-\mu^{\circ}) & \sum_{ijk}(T_{ijk}-\mu)(a_i+H_{ijk}) \\ \sum_{ijk}(T_{ijk}-\mu)(T^{\circ}_{ijk}-\mu^{\circ}) & \sum_{ijk}(T^{\circ}_{ijk}-\mu^{\circ})^2 & \sum_{ijk}(T^{\circ}_{ijk}-\mu^{\circ})(a_i+H_{ijk}) \\ \sum_{ijk}(T_{ijk}-\mu)(a_i+H_{ijk}) & \sum_{ijk}(T^{\circ}_{ijk}-\mu^{\circ})(a_i+H_{ijk}) & \sum_{ijk}(a_i+H_{ijk})^2 \end{bmatrix} .$$

The other first derivatives of L (A7.3) are given by

$$(A7.5)\quad \begin{pmatrix} \frac{\partial L}{\partial \mu} \\ \frac{\partial L}{\partial \mu^{\circ}} \end{pmatrix} = (\Sigma^{TT}, \Sigma^{TH}) \begin{pmatrix} T..-N..\mu \\ T^{\circ}..-N..\mu^{\circ} \\ \sum_i N_{i.} a_i + H.. \end{pmatrix}$$

and

$$(A7.6)\quad \frac{\partial L}{\partial a_i} = -(\Sigma^{TH'}, \Sigma^{HH}) \begin{pmatrix} T_{i.}-N_{i.}\mu \\ T^{\circ}_{i.}-N_{i.}\mu^{\circ} \\ N_{i.} a_i + H_{i.} \end{pmatrix}$$

Now if we set (A7.5) equal to $(0,0)'$, obtain a third equation $-\sum_i \partial L/\partial a_i = 0$ by using (A7.6), and pre-multiply the resulting set of three equations by Σ, then we will arrive at the solutions

$$\text{(A7.7)} \qquad \hat{\mu} = \bar{T}_{..} \quad , \quad \hat{\mu}^{\circ} = \bar{T}^{\circ}_{..}$$

for μ and μ°. By setting (A7.6) equal to 0 and solving for a_i, we get

$$\text{(A7.8)} \qquad a_i = -\bar{H}_{i.} - \frac{1}{\Sigma^{HH}} \Sigma^{TH'} \begin{pmatrix} \bar{T}_{i.} - \bar{T}_{..} \\ \bar{T}^{\circ}_{i.} - \bar{T}^{\circ}_{..} \end{pmatrix}$$

after substituting (A7.7). Next we plug (A7.7) and (A7.8) into (A7.4) and obtain the relations

$$\text{(A7.9)} \qquad N_{..}\Sigma_{TT} = W_1$$

and

$$\text{(A7.10)} \qquad N_{..}\Sigma_{TH} = \begin{pmatrix} \sum_i S_{THi.} \\ \sum_i S_{T^{\circ}Hi.} \end{pmatrix} - \frac{1}{\Sigma^{HH}} W_2 \Sigma^{TH} \quad ,$$

where W_1 and W_2 are as defined by (A5.5) and (A5.27) respectively. From (A7.2) and from the fact the $\Sigma\Sigma^{-1} = I$, it follows that

$$\Sigma_{TT} \Sigma^{TH} + \Sigma_{TH}\Sigma^{HH} = 0 \text{ (null matrix)} \quad ,$$

so that

$$\text{(A7.11)} \qquad N_{..}\Sigma_{TH} = - \frac{1}{\Sigma^{HH}} (N_{..}\Sigma_{TT})\Sigma^{TH} \quad .$$

Now we substitute (A7.11) and then (A7.9) into the left-hand side of (A7.10), and find that

$$\text{(A7.12)} \qquad - \frac{1}{\Sigma^{HH}} \Sigma^{TH} = (W_1 - W_2)^{-1} \begin{pmatrix} \sum_i S_{THi.} \\ \sum_i S_{T^{\circ}Hi.} \end{pmatrix}$$

upon solving for Σ^{TH}. Finally, we plug (A7.12) into (A7.8) and end up with

$$(A7.13)\quad \hat{a}_i = -\bar{H}_{i\cdot} + \left(\sum_{i'} S_{THi'\cdot}, \sum_{i'} S_{T^\circ Hi'\cdot}\right) \begin{bmatrix} \sum_{i'} S_{TTi'\cdot} & \sum_{i'} S_{TT^\circ i'\cdot} \\ \sum_{i'} S_{TT^\circ i'\cdot} & \sum_{i'} S_{T^\circ T^\circ i'\cdot} \end{bmatrix}^{-1} \begin{pmatrix} \bar{T}_{i\cdot} - \bar{T}_{\cdot\cdot} \\ \bar{T}^\circ_{i\cdot} - \bar{T}^\circ_{\cdot\cdot} \end{pmatrix}$$

as the maximum-likelihood estimator for the equating parameter a_i.

Note 8

The logarithm of the product over (i,j,k) of the expressions (9.2) is

$$(A8.1)\quad L = \text{const} + \sum_j N_{\cdot j} \log|\beta_j| - \tfrac{1}{2} \sum_{ijk} (\alpha_j + \beta_j C_{ijk} - \nu T_{ijk} - a_i - b_i H_{ijk})^2 \quad .$$

Taking derivatives, we find

$$(A8.2)\quad \frac{\partial L}{\partial a_i} = \sum_j N_{ij}\,\alpha_j + \sum_j \beta_j C_{ij} - \nu\, T_{i\cdot} - N_{i\cdot} a_i - b_i H_{i\cdot} \quad ,$$

$$(A8.3)\quad \frac{\partial L}{\partial b_i} = \sum_j \alpha_j H_{ij} + \sum_j \beta_j \sum_k C_{ijk} H_{ijk} - \nu \sum_{jk} T_{ijk} H_{ijk} - a_i H_{i\cdot} - b_i \sum_{jk} H^2_{ijk} \quad ,$$

$$(A8.4)\quad \frac{\partial L}{\partial \nu} = \sum_j \alpha_j T_{\cdot j} + \sum_j \beta_j \sum_{ik} C_{ijk} T_{ijk} - \nu \sum_{ijk} T^2_{ijk} - \sum_i a_i T_{i\cdot} - \sum_i b_i \sum_{jk} T_{ijk} H_{ijk} \quad ,$$

$$(A8.5)\quad \frac{\partial L}{\partial \alpha_j} = -N_{\cdot j}\alpha_j - \beta_j C_{\cdot j} + \nu\, T_{\cdot j} + \sum_i N_{ij} a_i + \sum_i b_i H_{ij} \quad ,$$

and

$$(A8.6)\quad \frac{\partial L}{\partial \beta_j} = (N_{\cdot j}/\beta_j) - \alpha_j C_{\cdot j} - \beta_j \sum_{ik} C^2_{ijk} + \nu \sum_{ik} C_{ijk} T_{ijk} + \sum_i a_i C_{ij} + \sum_i b_i \sum_k C_{ijk} H_{ijk} .$$

If we set (A8.2) equal to 0 and solve for a_i, we get (9.4). Next we set

(A8.3) equal to 0, make the substitution (9.4), and solve for b_i; this gives us (9.5). Now if we substitute first (9.4) and then (9.5) into (A8.4-A8.5) after setting the latter equal to 0, we wind up eventually with the system (9.12). For this system (9.12) we obtain a solution of the form (9.13), which will generally be a unique solution except for the fact that a constant can be added to each $\hat{\alpha}_j$ (see Note 9 below for details). After setting (A8.6) equal to 0, we plug in first (9.4) and then (9.5) to obtain

$$(A8.7a)\quad \sum_J \left[\delta_{jJ} \sum_{ik} C_{ijk}^2 - \sum_i \frac{C_{ij}C_{iJ}}{N_{i.}} - \sum_i \frac{(S_{CHij} + \bar{C}_{ij}d_{ij})(S_{CHiJ} + \bar{C}_{iJ}d_{iJ})}{S_{HHi.}} \right] \beta_J$$

$$- \sum \left[-\delta_{jJ} C_{.j} + \sum_i \frac{N_{iJ}C_{ij}}{N_{i.}} + \sum_i \frac{d_{iJ}(S_{CHij} + \bar{C}_{ij}d_{ij})}{S_{HHi.}} \right] \alpha_J$$

$$- \left[\sum_i (S_{CTij} + \bar{C}_{ij}e_{ij}) - \sum_i \frac{S_{THi.}(S_{CHij} + \bar{C}_{ij}d_{ij})}{S_{HHi.}} \right] \nu = N_{.j}/\beta_j \quad ,$$

i.e.,

$$(A8.7b)\quad \sum_J u_{jJ}\beta_J - \sum_J g_{Jj}\alpha_J - g_{\nu j}\nu = N_{.j}/\beta_j \quad ;$$

note that the system (A8.7) is unaffected if each α_J is altered by the same additive constant, inasmuch as $\sum_J g_{Jj} = 0$. Upon substituting (9.13) into (A8.7b), we end up with the system (9.17), the solution of which is discussed below (see Notes 10 and 11). Once the $\hat{\beta}_j$'s are determined, the estimates of the other parameters are of course obtained via formulas (9.13), (9.5), and (9.4).

Note 9

Up through the point of determining the estimates of the α_j's, ν, the a_i's, and the b_i's in terms of the β_j's, we are dealing with nothing more than a strictly linear analysis of variance model. Up to this point, the problem may be thought of as that of maximizing L (A8.1) for fixed values of the β_j's. This is equivalent to the problem of finding the least squares estimates of the a_i's, the b_i's, the α_j's, and ν under the linear model

$$(A9.1) \qquad E(\beta_j C_{ijk}) = -\alpha_j + \nu T_{ijk} + a_i + b_i H_{ijk} \quad .$$

Consequently, we may apply some of the broad theoretical results for the general linear model (as given, e.g., by Bose [3]) in examining certain facets of the equation system (9.12). For this purpose, we consider that the β_j's on the right-hand side of (9.12) and on the left-hand side of (A9.1) are fixed.

The model (A9.1) actually bears some resemblance to the model for the incomplete block design, or for the two-way layout with unequal numbers in the cells. Consequently, there is some similarity in the formulas, the normal equations, and the theoretical development. We first examine the matter of the rank of the matrix F (9.8). Now the rank of F cannot exceed n, by virtue of (9.14). By using methods akin to those used for incomplete block designs, we will develop a set of conditions for the rank of F to be exactly n. This set of conditions will be sufficient but not necessary; it should be adequate for practical purposes.

First we make the trivial assumption that ν is estimable [3] under the model (A9.1). A sufficient condition for this to be true is that there exist three students, identified by (i,j,k_1), (i,j,k_2), and (i,j,k_3), who are from the same high school and who go to the same college, and for whom

$$\begin{vmatrix} 1 & T_{ijk_1} & H_{ijk_1} \\ 1 & T_{ijk_2} & H_{ijk_2} \\ 1 & T_{ijk_3} & H_{ijk_3} \end{vmatrix} \neq 0 \quad .$$

Another assumption which we make, of course, is that $S_{HHi.} > 0$ for every high school.

Now it can be shown that the rank of F will be as high as n if and only if the contrast $(\alpha_J - \alpha_j)$ is estimable for all pairs (j,J). We now give a sufficient set of conditions for all such contrasts to be estimable. Let m_o $(\leq m)$ be the number of different high schools (i-values) such that, for each of these m_o i's, there exists at least one j-value (college) such that $N_{ij} > 1$ and such that $H_{ijl}, H_{ij2}, \ldots, H_{ijN_{ij}}$ are not all alike. (This implies that b_i is estimable for each of these m_o high schools.) Consider the $m_o \times n$ matrix whose general element is N_{ij}, with the rows of the matrix referring to the m_o specified high schools and the columns to the totality of the n colleges. If this $m_o \times n$ matrix constitutes the incidence matrix for a <u>connected</u> incomplete block design [i.e., if, for every pair (j,J), there exists a connecting chain $N_{i_1j}, N_{i_1j_1}, N_{i_2j_1}, N_{i_2j_2}, N_{i_3j_2}, \ldots,$ $N_{i_{r-1}j_{r-1}}, N_{i_rj_{r-1}}, N_{i_rJ}$, all of whose elements come from the matrix and are > 0], then $(\alpha_J - \alpha_j)$ will be estimable for all pairs of colleges (j,J), and, consequently, F will be of rank n. We have not supplied all of the fine details in the argument we have used in this paragraph, but the argument bears some resemblance to certain standard developments in incomplete block design theory.

Thus we have a sufficient condition for F to be of rank n. The condition would not appear to be too difficult to check for. Offhand, there

would appear to be little doubt of the condition being satisfied under a large central prediction system.

If F were not of rank n, then not all differences $(\alpha_J - \alpha_j)$ would be estimable, which would mean there would be no basis for comparing grades from certain high schools. The calculations for estimating the parameters would then become a bit more complicated. However, we will not consider further the case where F has rank smaller than n, since such a case should not arise in practice with a large system, and the sufficient condition of the previous paragraph ought to be satisfied without any trouble. From now on, we will assume that F is of rank n.

If $(\alpha_1, \alpha_2, \ldots, \alpha_n, \nu)'$ and $(\alpha_1', \alpha_2', \ldots, \alpha_n', \nu')'$ represent two different solutions of (9.12), then it follows that $\nu = \nu'$ and that the difference $\alpha_j - \alpha_j'$ is the same for all j. To show this, we consider the two equations (9.12), one with $(\alpha_1, \alpha_2, \ldots, \alpha_n, \nu)'$ and the other with $(\alpha_1', \alpha_2', \ldots, \alpha_n', \nu')'$. We subtract the latter from the former, and obtain the null vector on the right-hand side and F times $(\alpha_1 - \alpha_1', \alpha_2 - \alpha_2', \ldots, \alpha_n - \alpha_n', \nu - \nu')'$ on the left-hand side. Thus $(\alpha_1 - \alpha_1', \alpha_2 - \alpha_2', \ldots, \alpha_n - \alpha_n', \nu - \nu')$ lies in the vector space orthogonal to the rows (or columns) of F. But, since F is of rank n, the vector space orthogonal to the rows of F must be of rank 1, and in fact must have as its basis the vector $(1, 1, \ldots, 1, 0)'$ in view of (9.14). This is sufficient to complete the proof, and we conclude that the solution of (9.12) is unique except for the fact that the α_j's may all be altered by the same additive constant.

Now (9.13) will be a solution of (9.12) if F^* is any conditional inverse [3] of F. More particularly, we will indicate here how to obtain a specific $F^*([n+1] \times [n+1])$ which can satisfactorily be used. We consider the matter simply in terms of finding a solution of the system (9.12). By

virtue of (9.14), one of the first n rows of (9.12) is superfluous; accordingly, we arbitrarily eliminate the first row of the system (9.12). Next, we arbitrarily decide to pick the solution for which $\alpha_1 = 0$; accordingly, we may knock out the first column of what remains of F, and at the same time remove α_1 from the vector. We are left with n equations in n unknowns, which will have a unique solution. What remains of F is the nxn matrix in the lower right-hand corner of F, which we call $F_{11}(nxn)$. From (9.14) and from the fact that F is of rank n, it follows that F_{11} is of rank n. Hence F_{11}^{-1} (nxn) exists. We may thus take F*[(n+1)x(n+1)] to be a matrix containing F_{11}^{-1} in its lower right-hand corner and zeroes elsewhere. It may also be verified directly that, with such an F*, (9.13) satisfies (9.12).

Thus the problem of finding this F* is essentially the problem of inverting F_{11}. $F_{11}(nxn)$ is of course a huge matrix to invert. Note from (9.8,9.9,9.14), however, that (except for the last row and last column) the diagonal elements of F_{11} will be large and positive while the off-diagonal elements will be small and apparently nearly all negative. Because of this, F_{11} should be much easier to invert than would otherwise be the case if the main diagonal elements were not relatively large. As for the last row and last column of F_{11}, it might possibly help, in inverting the matrix, to partition off this row and column and then utilize formulas similar to (A4.8).

Note 10

We prove the uniqueness of any solution of (9.17) such that all β_j's are > 0. We note first that this system (9.17) can be arrived at in a different manner. In line with the discussion of Notes 8 and 9, let us observe first that, if (9.4), (9.5), and (9.13) are plugged into (A8.1), this will result in maximizing L (A8.1) for any fixed values of the β_j's. After

these substitutions, the expression within the parentheses in (A8.1) will become a linear function of the β_j's, of the form $\sum_J z_{ijkJ}\beta_J$, say, and L (A8.1) itself will become

$$(A10.1) \qquad L = \text{const} + \sum_j N_{\cdot j}\log|\beta_j| - \tfrac{1}{2}\sum_{ijk}\sum_J\Big(\sum z_{ijkJ}\beta_J\Big)^2$$

$$= \text{const} + \sum_j N_{\cdot j}\log|\beta_j| - \tfrac{1}{2}\underline{\beta}'Z'Z\underline{\beta} \quad ,$$

where the matrix $Z(N.. \times n)$ contains the general element z_{ijkJ} in its (i,j,k)-th row and J-th column. To find the values of the β_j's which maximize (A10.1), we differentiate (A10.1) and set the derivatives equal to 0:

$$(A10.2) \qquad \frac{\partial L}{\partial \underline{\beta}} = D_\beta^{-1}\underline{N} - Z'Z\,\underline{\beta} = \underline{0} \qquad \text{(null vector)} \quad .$$

Now the system (A10.2) must necessarily be the same as the system (9.17); thus

$$(A10.3) \qquad Z'Z = A \quad .$$

We next use (A10.2) to find the matrix of second derivatives of L (A10.1):

$$(A10.4) \qquad \frac{\partial^2 L}{\partial \underline{\beta}^2} = -D_{N/\beta^2} - Z'Z \quad ,$$

where $D_{N/\beta^2}(n \times n)$ denotes a diagonal matrix with the elements $N_{\cdot j}/\beta_j^2$ along the main diagonal. Now this matrix (A10.4) is clearly negative definite for all values of the β_j's. Furthermore, L(A10.1) and its derivatives are continuous throughout the set of points for which

$$(A10.5) \qquad \beta_j > 0 \quad \text{for all } j \quad .$$

Thus, by using the same line of argument that was set forth in the next to the last paragraph of Note 3, we conclude that there can be no more than one solution of (A10.2) which lies in the set (A10.5). Hence, since (9.17) is

the same system as (A10.2), any solution $\underline{\beta}$ of (9.17) which satisfies (A10.5) must necessarily be unique in the set (A10.5).

Note 11

The details of selecting a technique for solving the system (9.17) would best be left to a specialist in computers and numerical analysis. However, we will indicate here some avenues of approach which may be promising.

First we consider the generalized Newton-Raphson method (see, e.g., [7, p.203 ff.], [6, p. 135 ff.], or [10, p. 171 ff.]). In general, if we are trying to solve a system of n simultaneous equations in n unknowns, of the form

$$\text{(A11.1a)} \qquad \phi^{(j)}(\beta_1, \beta_2, \ldots, \beta_n) = 0 \qquad (j = 1,2, \ldots, n) \quad ,$$

i.e.,

$$\text{(A11.1b)} \qquad \underline{\phi}(\underline{\beta}) = \underline{0} \quad ,$$

where $\underline{\phi}$ is an $(n \times 1)$ vector function, then the generalized Newton-Raphson method uses the iteration formula

$$\text{(A11.2)} \qquad \underline{\beta}_{new} = \underline{\beta}_{old} - [\phi_1(\underline{\beta}_{old})]^{-1} \; \underline{\phi}(\underline{\beta}_{old}) \quad ,$$

where $[\phi_1(\underline{\beta})]$ denotes an $(n \times n)$ matrix whose general element in the j'-th row and j-th column is $\partial\phi^{(j')}/\partial\beta_j$. Sometimes $[\phi_1(\underline{\beta}_o)]^{-1}$, where $\underline{\beta}_o$ denotes the value of $\underline{\beta}$ in the initial iteration, may be used in place of $[\phi_1(\underline{\beta}_{old})]^{-1}$ in (A11.2) (see, e.g., [10, p.172]); this spares us from having to invert an $n \times n$ matrix at each iteration, and hence only a single $n \times n$ matrix has to be inverted.

We turn our attention to the specific system (9.17). Taking

(A11.3) $$\underline{\phi}(\underline{\beta}) = D_{\beta}^{-1}\,\underline{N} - A\,\underline{\beta} \quad ,$$

we find

(A11.4) $$\phi_1(\underline{\beta}) = -\,(D_{N/\beta^2} + A) \quad .$$

Thus we obtain our iteration formula by substituting (A11.3) and (A11.4) into (A11.2)

Alternatively, we could take $\underline{\phi}$ to be

(A11.5) $$\underline{\phi}(\underline{\beta}) = \underline{N} - D_{\beta}A\,\underline{\beta} \quad ,$$

for which

(A11.6) $$\phi_1(\underline{\beta}) = -(D_{\beta}A + D_{A\underline{\beta}}) \quad ,$$

where $D_{A\underline{\beta}}$ denotes a diagonal matrix whose diagonal elements are the elements of the vector $(A\underline{\beta})$. Offhand, though, this formulation (A11.5) would not appear to offer any advantage over (A11.3).

The other method which we will consider for solving (9.17) is the method of steepest descent (see, e.g., [4], [6, p.132 ff.], or [10, p. 175 ff.]). Strictly speaking, the method of steepest descent is used for finding the point at which a function of n variables assumes a realtive minimum, rather than for solving a system of the form (A11.1); however, this latter problem can be handled as a special case (see [6, pp. 132-133]). Let $\psi(\underline{\beta})$ denote the function which is to be minimized. Let $\underline{\psi}_1(\underline{\beta})$ denote an $nx1$ vector whose j-th element is $\partial\psi/\partial\beta_j$. Then the iteration formula for the method of steepest descent is

(A11.7) $$\underline{\beta}_{new} = \underline{\beta}_{old} - \lambda(\underline{\beta}_{old})\,\underline{\psi}_1(\underline{\beta}_{old}) \quad ,$$

where $\lambda(\underline{\beta})$ is a scalar whose determination we now consider. Ideally,

$\lambda(\underline{\beta})$ should be (see [10, p.176, equation (6.27)] or [4, p.260, equation (5)]) the smallest positive root of the equation

(A11.8) $\qquad \gamma'(\lambda) = 0 \qquad ,$

where

(A11.9) $\qquad \gamma(\lambda) = \psi(\underline{\beta} - \lambda \underline{\psi}_1(\underline{\beta})) \quad .$

However, for use in (A11.7) it is adequate to obtain an approximation to this ideal value of λ. In cases where the actual minimum value of ψ is 0, we can get the Newton-Raphson first approximation to the (supposed) root of $\gamma(\lambda) = 0$ rather than to the root of (A11.8), so that the formula

(A11.10) $\qquad \lambda(\underline{\beta}) = \psi(\underline{\beta}) / \underline{\psi}_1'(\underline{\beta})\, \underline{\psi}_1(\underline{\beta})$

should give a satisfactory value for λ (see [10, p.176, formula (6.28)] or [4, p.259, last paragraph]). However, in cases where the minimum value of ψ is not 0, the formula (A11.10) may lead to trouble (see [4, p.259, footnote 7]), and it would appear to be better (if feasible) to use the Newton-Raphson first approximation to the solution of (A11.8) itself, which is

(A11.11) $\qquad \lambda(\underline{\beta}) = \underline{\psi}_1'(\underline{\beta})\underline{\psi}_1(\underline{\beta}) / \underline{\psi}_1'(\underline{\beta})\psi_2(\underline{\beta})\underline{\psi}_1(\underline{\beta}) \qquad ,$

where $\psi_2(\underline{\beta})$ is the $n \times n$ matrix of the second derivatives of $\psi(\underline{\beta})$.

We now consider how the method of steepest descent can be applied to the system (9.17). If we take

(A11.12) $\qquad \psi(\underline{\beta}) = -\sum_j N_{\cdot j} \log|\beta_j| + \tfrac{1}{2}\underline{\beta}'A\underline{\beta} \qquad ,$

then

(A11.13) $$\underline{\Psi}_1(\underline{\beta}) = -D_{\beta}^{-1}\underline{N} + A\,\underline{\beta} \quad ,$$

so that the problem of solving (9.17) is equivalent to the problem of finding a point at which $\psi(\underline{\beta})$ (A11.12) assumes a minimum. Thus we plug (A11.13) into (A11.7). For $\lambda(\underline{\beta})$ we would probably use (A11.11), in which case we need the matrix

(A11.14) $$\Psi_2(\underline{\beta}) = D_{N/\beta^2} + A \quad .$$

There are also other ways in which the method of steepest descent can be applied for solving (9.17). For instance, let us set

(A11.15) $$\psi(\underline{\beta}) = \underline{\phi}'(\underline{\beta})\,\underline{\phi}\,(\underline{\beta}) \quad ,$$

where $\underline{\phi}(\underline{\beta})$ is as given by (A11.3). Then $\psi(\underline{\beta})$ (A11.15) assumes its minimum value (of 0) if and only if $\underline{\beta}$ satisfies (9.17). Thus we use the method of steepest descent to find the point where $\psi(\underline{\beta})$ (A11.15) is minimal. From (A11.15) and (A11.3) we get

(A11.16) $$\psi(\underline{\beta}) = (D_{\beta}^{-1}\underline{N} - A\,\underline{\beta})'\,(D_{\beta}^{-1}\underline{N} - A\underline{\beta}) \quad .$$

Hence

(A11.17) $$\underline{\Psi}_1(\underline{\beta}) = 2(A + D_{N/\beta^2})(A\underline{\beta} - D_{\beta}^{-1}\underline{N}) \quad ,$$

and we plug (A11.17) into (A11.7). For $\lambda(\underline{\beta})$ this time we should probably use (A11.10), for which we need only (A11.16) and (A11.17).

Notice that no matrix inversions are required in either of the iteration procedures which are based on the method of steepest descent. This is in contrast to our two iteration procedures based on the generalized Newton-Raphson method, both of which require matrix inversion; however, in (A11.2) we could

use an approximation to the inverse of $\phi_1(\beta)$ rather than calculate the exact inverse.

It is possible that still other techniques for solving (9.17) should be considered; our treatment here is not intended to be exhaustive. For references to other methods for solving a system of n equations in n unknowns, see, e.g., [8, p.215 ff.] and [10, Chapter 6].

In order to utilize any iterative procedure for solving (9.17), it is necessary to have an initial value of β (which we call β_o) to start off with. Of course, the "closer" β_o is to the true solution of (9.17), the better off we should be. What we propose is to use for β_o the maximum-likelihood estimates of the β_j's under the C|T model which was covered in Section 5. Thus the elements of β_o could be calculated from (5.7) after (5.6) has been solved for ν. Such a β_o should not be violently different from the exact solution of (9.17) which we are aiming toward. Incidentally, it would appear that we should throw out any college for which $S_{CT.j}$ is < 0 [this causes (5.7) to be negative also], but it would not seem to be too likely that such a condition would ever crop up in the first place.

An alternative possibility for choosing β_o would be to use the estimates of the β_j's which are obtained by going through both steps of the second approach as described in Section 7. A β_o so chosen might in many cases be "closer" to the solution of (9.17) than the β_o which was described in the previous paragraph, but the additional computational labor which would be required might or might not be worth it.

It is hoped that the material presented here in Note 11 will provide an adequate foundation for finding a means of solving (9.17) at a reasonable cost.

Note 12

Under this simplified model in which we omit b_i' in the formula for h_{ijk} (7.3), the distribution of C_{ijk} given T_{ijk} and H_{ijk} is of the form

$$(A12.1) \qquad (2\pi)^{-\frac{1}{2}}|\beta_j|\, e^{-\frac{1}{2}(\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-a_i-bH_{ijk})^2} \quad .$$

The theoretical development for this model (A12.1) will be quite similar to that for the model (9.2). The logarithm of the product over (i,j,k) of the expressions (A12.1) is the same as (A8.1), except with b_i replaced by b. Let us use (A8.1°),(A8.2°), (A8.4°), (A8.5°), and (A8.6°) to designate the same equations as (A8.1), (A8.2), (A8.4), (A8.5), and (A8.6) respectively, except with b_i replaced by b. Then the partial derivatives of L(A8.1°) with respect to a_i, ν, α_j, and β_j are given respectively by (A8.2°), (A8.4°), (A8.5°), and (A8.6°). Also we find from (A8.1°) that

$$(A12.2) \qquad \frac{\partial L}{\partial b} = \sum_j \alpha_j H_{.j} + \sum_j \beta_j \sum\sum_{ik} C_{ijk}H_{ijk} - \nu \sum\sum\sum_{ijk} T_{ijk}H_{ijk} - \sum_i a_i H_{i.} - b \sum\sum\sum_{ijk} H_{ijk}^2 \quad .$$

Upon setting (A8.2°) equal to 0 and solving for a_i, we find

$$(A12.3) \qquad \hat{a}_i = (1/N_{i.})\Big(\sum_j N_{ij}\hat{\alpha}_j + \sum_j \hat{\beta}_j C_{ij}\Big) - \hat{\nu}\bar{T}_{i.} - \hat{b}\bar{H}_{i.} \quad .$$

Next we set (A8.5°), (A8.4°), and (A12.2) equal to 0 and make the substitution (A12.3). Then, after the terms in the β_j's are isolated on the right-hand side, this set of equations becomes

$$(A12.4) \qquad F_o \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \nu \\ b \end{pmatrix} = G_o \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \quad ,$$

where F_o ([n+2] x [n+2]) and G_o ([n+2]xn) are defined as follows. The general element of G_o in the j-th row and J-th column is

$$(A12.5) \qquad g_{ojJ} = -\delta_{jJ}C_{.j} + \sum_i \frac{N_{ij}C_{iJ}}{N_{i.}}$$

for the first n rows,

$$(A12.6) \qquad g_{o\nu J} = \sum_i (S_{CTiJ} + \bar{C}_{iJ}e_{iJ})$$

for the (n+1)-th row, and

$$(A12.7) \qquad g_{obJ} = \sum_i (S_{CHiJ} + \bar{C}_{iJ}d_{iJ})$$

for the (n+2)-th row [e_{iJ} and d_{iJ} being defined by (9.3)]. F_o is symmetric and has the form

$$(A12.8) \qquad F_o = \left[\begin{array}{ccc|cc} f_{o11} & f_{o12} & \cdots & f_{o1n} & f_{o1\nu} & f_{o1b} \\ f_{o21} & f_{o22} & \cdots & f_{o2n} & f_{o2\nu} & f_{o2b} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ f_{on1} & f_{on2} & \cdots & f_{onn} & f_{on\nu} & f_{onb} \\ \hline f_{o\nu 1} & f_{o\nu 2} & \cdots & f_{o\nu n} & f_{o\nu\nu} & f_{o\nu b} \\ f_{ob1} & f_{ob2} & \cdots & f_{obn} & f_{ob\nu} & f_{obb} \end{array}\right],$$

where

$$(A12.9) \qquad f_{ojJ} = \delta_{jJ}N_{.j} - \sum_i \frac{N_{ij}N_{iJ}}{N_{i.}} \quad ,$$

$$(A12.10) \qquad f_{oj\nu} = -T_{.j} + \sum_i N_{ij}\bar{T}_{i.} \qquad (=f_{o\nu j}) \quad ,$$

$$(A12.11) \qquad f_{ojb} = -H_{.j} + \sum_i N_{ij}\bar{H}_{i.} \qquad (=f_{obj}) \quad ,$$

$$(A12.12) \qquad f_{o\nu\nu} = \sum_i S_{TTi.} \quad ,$$

(A12.13) $$f_{o\nu b} = \sum_i S_{THi.} \qquad (=f_{ob\nu}) \quad ,$$

and

(A12.14) $$f_{obb} = \sum_i S_{HHi.} \quad .$$

If we make the trivial assumption that ν and b are estimable under the linear model with fixed β_j's which is analogous to (A9.1), then it follows immediately (by appealing to incomplete block design theory) that F_o (A12.8) is of rank n , so long as the mxn matrix of the N_{ij}'s constitutes the incidence matrix of a connected incomplete block design (see Note 9 for the definition of "connected"). By using virtually the same technique which was indicated in the latter part of Note 9 for determining F*, we can obtain a matrix F_o^* such that

(A12.15) $$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_n \\ \hat{\nu} \\ \hat{b} \end{pmatrix} = F_o^* G_o \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix}$$

is a solution of (A12.4) for the α_j's, ν, and b in terms of the β_j's. (See Note 9 for certain remarks which apply also in the present development.) At this point it remains only to solve for the β_j's: we set (A8.6°) equal to 0, plug in first (A12.3) and then (A12.15), and wind up finally with the system

(A12.16) $$A_o \underline{\beta} = D_\beta^{-1} \underline{N} \quad ,$$

where

(A12.17) $$A_o(nxn) = U_o - G_o' F_o^* G_o$$

and U_o in (A12.17) is an nxn matrix whose general element is

$$(A12.18) \qquad u_{ojJ} = \delta_{jJ} \sum_i \sum_k C_{ijk}^2 - \sum_i \frac{C_{ij}C_{iJ}}{N_{i.}} \quad .$$

Since the system (A12.16) is of the same form as (9.17), we may refer at this point to Notes 10 and 11 for pertinent information concerning the solution of (A12.16).

Thus we must first obtain F_o^* and A_o, then solve (A12.16) to get the $\hat{\beta}_j$'s, and finally obtain the estimates of the remaining parameters via (A12.15) and (A12.3). As we can see, it turns out that the maximum-likelihood estimation procedure is formally almost the same for the model (A12.1) as for the model (9.2). Although many of the formulas are somewhat simpler under the model (A12.1) [compare (A12.5), (A12.9), and (A12.18) with their counterparts, e.g.], the computational labor required for the two most difficult parts of the calculations [i.e., obtaining F_o^* or F^*, and solving (A12.16) or (9.17)] does not appear to be reduced at all by simplifying the model so as to eliminate the b_i''s.

Note 13

If the model (9.2) is simplified by eliminating the β_j's, then the estimation procedure is exactly the same as that indicated in Section 9, except that we stop just before (9.15), and we replace with 1's all $\hat{\beta}_j$'s or β_j's which appear in formulas (9.4), (9.5), (9.12), and (9.13). Note that we no longer have to solve the non-linear system (9.17), so that the computational burden is reduced considerably by using a simplified model from which the β_j's have been eliminated.

The distribution of C_{ijk} given T_{ijk} and H_{ijk} under this simplified model

is, of course, of the form

$$(A13.1) \qquad (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(C_{ijk}+\alpha_j-\nu T_{ijk}-a_i-b_i H_{ijk})^2/\sigma^2} .$$

Thus this model (A13.1) is actually a linear model; in fact, it is essentially the same thing as (A9.1).

If the model (A12.1) is simplified by eliminating the β_j's, then the development proceeds exactly as in Note 12, except that we stop just before (A12.16), and we replace with 1's all $\hat{\beta}_j$'s or β_j's which appear in formulas (A12.3), (A12.4), and (A12.15).

Note 14

If the model (9.2) is altered to include more than one T-variable, then there will be no substantial change either in the theoretical development or in the amount of computational labor required to obtain the estimates. We treat the case of exactly two T-variables, since this will be sufficient to indicate what happens with a general number of T-variables. The conditional distribution of C_{ijk} given T_{ijk}, T°_{ijk}, and H_{ijk} is then

$$(A14.1) \qquad (2\pi)^{-\frac{1}{2}} |\beta_j| e^{-\frac{1}{2}(\alpha_j+\beta_j C_{ijk}-\nu T_{ijk}-\nu^{\circ} T^{\circ}_{ijk}-a_i-b_i H_{ijk})^2} .$$

The manipulations that are involved in getting the formulas for the estimates under the model (A14.1) are practically the same as those outlined in Note 8 above. What we will do here will be simply to indicate how the various computing formulas of Section 9 are altered when there are two T-variables instead of one:

(i) In the formula for $\hat{a}_i$(9.4), we include an additional term $(-\hat{\nu}^{\circ}\bar{T}^{\circ}_{i.})$.

(ii) In the formula for $\hat{b}_i$(9.5), we include an additional term $(-\hat{\nu}^{\circ}S_{T^{\circ}Hi.})$

inside the square brackets.

(iii) In the matrix G, we add an extra row, so that G will be (n+2)x n. The elements of this (n+2)-th row are specified by the formula

$$(A14.2)\quad g_{\nu^\circ J} = \sum_i (S_{CT^\circ iJ} + \bar{C}_{iJ} e^\circ_{iJ}) - \sum_i \frac{S_{T^\circ Hi.}(S_{CHiJ} + \bar{C}_{iJ} d_{iJ})}{S_{HHi.}} \quad ,$$

where $e^\circ_{ij} = T^\circ_{ij} - N_{ij}\bar{T}^\circ_{i.}$ The first (n+1) rows of G are the same as before (9.6-9.7).

(iv) In the matrix F (9.8), we add another row and another column, so that F will be (n+2) x (n+2) and still symmetric. The additional elements of F are specified by

$$(A14.3)\quad f_{j\nu^\circ} = -T^\circ_{.j} + \sum_i N_{ij}\bar{T}^\circ_{i.} + \sum_i \frac{d_{ij} S_{T^\circ Hi.}}{S_{HHi.}} \qquad (= f_{\nu^\circ j}) \quad ,$$

$$(A14.4)\quad f_{\nu\nu^\circ} = \sum_i \left(S_{TT^\circ i.} - \frac{S_{THi.} S_{T^\circ Hi.}}{S_{HHi.}}\right) \qquad (= f_{\nu^\circ\nu}) \quad ,$$

and

$$(A14.5)\quad f_{\nu^\circ\nu^\circ} = \sum_i \left(S_{T^\circ T^\circ i.} - \frac{S^2_{T^\circ Hi.}}{S_{HHi.}}\right) \quad .$$

The original elements of F are the same as before (9.9-9.11).

(v) At the bottom of the vectors appearing on the left-hand sides of (9.12) and (9.13), we include a ν° and a $\hat{\nu}^\circ$ respectively. The matrix F* in (9.13) is now (n+2) x (n+2), of course. No change is made in formulas (9.15-9.17).

If it is desired to use more than one T-variable in any of the models covered by Note 12 and Note 13 above, then the various relevant formulas can be altered in an obvious manner to accommodate the additional T-variable(s).

As was the case with the model (9.2), the relative increase in computational labor which results from the additional T-variable(s) will not be substantial.

Note 15

We suppose that c_{ijk}(equated college grade), T_{ijk}, and h_{ijk}(equated high school grade) follow a trivariate normal distribution in the unselected population. Let $(\mu_c, \mu_T, \mu_h)'$ denote the mean vector and

$$\text{(A15.1)} \qquad \begin{bmatrix} \sigma_c^2 & \rho_{CT}\sigma_c\sigma_T & \rho_{CH}\sigma_c\sigma_h \\ \rho_{CT}\sigma_c\sigma_T & \sigma_T^2 & \rho_{TH}\sigma_T\sigma_h \\ \rho_{CH}\sigma_c\sigma_h & \rho_{TH}\sigma_T\sigma_h & \sigma_h^2 \end{bmatrix}$$

the variance matrix of this distribution. (Note that ρ_{CT} and ρ_{cT}, ρ_{CH} and and ρ_{ch}, and ρ_{TH} and ρ_{Th} can be used interchangeably.) By appealing (e.g.) to [1, p.29, Theorem 2.5.1], utilizing (A15.1), and finally making the substitution (7.3), we find that the conditional distribution of T_{ijk} given H_{ijk} has mean

$$\text{(A15.2)} \qquad E(T_{ijk}|H_{ijk}) = \mu_T + \rho_{TH}\frac{\sigma_T}{\sigma_h}(a_i' + b_i'H_{ijk} - \mu_h)$$

and variance

$$\text{(A15.3)} \qquad \text{var}(T_{ijk}|H_{ijk}) = \sigma_T^2(1-\rho_{TH}^2) \quad ,$$

and the conditional distribution of c_{ijk} given T_{ijk} and H_{ijk} has mean

$$\text{(A15.4)} \qquad E(c_{ijk}|T_{ijk},H_{ijk}) = \mu_c + \frac{\sigma_c(\rho_{CT}-\rho_{CH}\rho_{TH})}{\sigma_T(1-\rho_{TH}^2)}(T_{ijk}-\mu_T) + \frac{\sigma_c(\rho_{CH}-\rho_{CT}\rho_{TH})}{\sigma_h(1-\rho_{TH}^2)}(a_i'+b_i'H_{ijk}-\mu_h)$$

and variance

$$(A15.5)\quad \mathrm{var}(c_{ijk}|T_{ijk},H_{ijk}) = \frac{\sigma_c^2(1-\rho_{CT}^2-\rho_{CH}^2-\rho_{TH}^2+2\rho_{CT}\rho_{CH}\rho_{TH})}{1-\rho_{TH}^2} \quad .$$

[Incidentally, (A15.5) is also equal to 1 the way the model (9.2) is set up, but this fact need not concern us here.]

Now the second approach is based on the model (7.4), which we repeat here for convenience, but we append an asterisk to a_i and b_i:

$$(A15.6)\quad E(T_{ijk}|H_{ijk}) = a_i^* + b_i^* H_{ijk} \quad .$$

The third approach, strictly speaking, is based on the model (9.2), but, as we indicated in Section 10, we will figure the variances for the third approach on the basis of a simplified model which assumes that we have the exact rather than just the estimated values of the α_j's and β_j's in (9.2). This simplified model, which is of the form

$$(A15.7)\quad E(c_{ijk}|T_{ijk},H_{ijk}) = \nu T_{ijk}+a_i+b_i H_{ijk} \quad ,$$

is linear, of course, whereas (9.2) is not; thus the determination of the variances is facilitated considerably. Although the variances of the estimates of the parameters as figured on the basis of the model (A15.7) will presumably be slightly smaller than the true variances based on (9.2), the difference apparently should not be great, because the largeness of the $N_{\cdot j}$'s should result in the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s being relatively quite close to the α_j's and β_j's respectively.

Now we can use results from the theory of the general linear model to obtain the variances of the estimators under (A15.6) and (A15.7). We arrive ultimately at the formulas

$$(A15.8) \quad \mathrm{var}(\hat{a}_i^* + \hat{b}_i^* H) = \left[\frac{1}{N'_{i.}} + \frac{(H-\bar{H}'_{i.})^2}{S'_{HHi.}}\right] \mathrm{var}\,(T_{ijk}|H_{ijk})$$

for the special form of the second approach, and

$$(A15.9) \quad \mathrm{var}(\hat{a}_i + \hat{b}_i H) = \left[\frac{1}{N_{i.}} + \frac{(H-\bar{H}_{i.})^2}{S_{HHi.}} + (\text{terms pertaining to } \nu)\right] \mathrm{var}(c_{ijk}|T_{ijk}, H_{ijk})$$

for the third approach. The "terms pertaining to ν" in (A15.9) relate to the discrepancy between ν and $\hat{\nu}$, and can be ignored for practical purposes.

Comparing (A15.6) with (A15.2) and (A15.7) with (A15.4), we see that the a_i^* and b_i^* of (A15.6) are not the same thing as the a_i and b_i of (A15.7), since $(a'_i + b'_i H_{ijk})$ is multiplied by a factor

$$(A15.10) \quad b'^* = \rho_{TH}\sigma_T/\sigma_h$$

in (A15.2), and by a factor

$$(A15.11) \quad b' = \sigma_c(\rho_{CH}-\rho_{CT}\rho_{TH})/\sigma_h(1-\rho_{TH}^2)$$

in (A15.4). In order to make (A15.8) comparable with (A15.9), we multiply $(\hat{a}_i^* + \hat{b}_i^* H)$ in (A15.8) by the ratio of (A15.11) to (A15.10):

$$(A15.12) \quad \mathrm{var}\left[\frac{b'}{b'^*}(\hat{a}_i^* + \hat{b}_i^* H)\right] = \left[\frac{1}{N'_{i.}} + \frac{(H-\bar{H}'_{i.})^2}{S'_{HHi.}}\right] \frac{\sigma_c^2(\rho_{CH}-\rho_{CT}\rho_{TH})^2}{\sigma_T^2\rho_{TH}^2(1-\rho_{TH}^2)^2} \mathrm{var}\,(T_{ijk}|H_{ijk})\,.$$

Finally, in order to obtain (10.3), we divide (A15.9) (with the "terms pertaining to ν" omitted) by (A15.12), after first substituting (A15.3) into (A15.12) and (A15.5) into (A15.9).

Note 16

Since the conditional expectation of (11.4) is only negligibly different from 0, the conditional variance of (11.4) is essentially

$$(A16.1)\quad E\left[(c_{ijk}-\hat{c}_{ijk})^2 \,|\, T_{ijk},H_{ijk}\right] = E\left[\{c_{ijk}-E(c_{ijk}|T_{ijk},H_{ijk})\}^2 \,|\, T_{ijk},H_{ijk}\right]$$
$$+2E\left[\{c_{ijk}-E(c_{ijk}|T_{ijk},H_{ijk})\}\{E(c_{ijk}|T_{ijk},H_{ijk})-\hat{c}_{ijk}\}\,|\,T_{ijk},H_{ijk}\right]$$
$$+E\left[\{E(c_{ijk}|T_{ijk},H_{ijk})-\hat{c}_{ijk}\}^2\,|\,T_{ijk},H_{ijk}\right] \quad .$$

We consider individually the three terms on the right-hand side of (A16.1). The first term is the same thing as (A15.5), which is equal to 1 under either of the formulations (9.2) or (7.18). The second term is 0, inasmuch as c_{ijk} and $\hat{c}_{ijk}$ will be independent (in the conditional distribution, at least) no matter what technique was used for obtaining the $\hat{a}_i$'s and $\hat{b}_i$'s. In evaluating the third term, we plug in (11.3) and (11.2), and first of all we simplify matters by putting $\hat{\nu} = \nu$ (an approximation which should make no practical difference). Then what we have is

$$(A16.2)\quad E\left[\{(\hat{a}_i+\hat{b}_iH_{ijk})-(a_i+b_iH_{ijk})\}^2 \,|\, T_{ijk},H_{ijk}\right] \quad ,$$

which under the third approach is essentially (A15.9), i.e., essentially

$$(A16.3)\quad \frac{1}{N_{(i).}} + \frac{(H_{ijk}-\bar{H}_{(i).})^2}{S_{HH(i).}}$$

since $\mathrm{var}(c_{ijk}|T_{ijk},H_{ijk}) = 1$. Adding 1 [the first term on the right-hand side of (A16.1)] to (A16.3), we get (11.5).

In considering (A16.2) for the special form of the second approach, we will treat only the case where the $\hat{a}_i$'s and $\hat{b}_i$'s of (7.5) and (7.6) are

obtained from a set of (T,H) data which is different from that for which the c_{ijk}'s are to be predicted. (If the two sets of data are the same, then the formulas which follow may be somewhat off, particularly so if $N_{i.}$ is small.) We can thus assume that neither $\hat{a}_i$ nor $\hat{b}_i$ is a function of T_{ijk} or H_{ijk}. Hence (A16.2) will be approximately the same thing as (A15.12), which can also be written as

$$(A16.4) \quad \left[\frac{1}{N'_{i.}} + \frac{(H-\bar{H}'_{i.})^2}{S'_{HHi.}}\right](\nu^\circ)^2 \ \mathrm{var}(T|H) \quad .$$

The ν° in (A16.4) is of course the parameter which appears in (7.18); var $(T|H)$ is the variance under the model (7.4), and of course does not depend on H [see also (A15.3)]. We can assume that ν° and var$(T|H)$ can both be estimated with negligible error.

In (A16.4) we write H_{ijk} instead of H, and we put parentheses around the i's everywhere else to indicate reference to a different set of data. After then adding 1, we end up with

$$(A16.5) \quad \sigma^2_{(c-\hat{c})} = 1 + \left[\frac{1}{N'_{(i).}} + \frac{(H_{ijk}-\bar{H}'_{(i).})^2}{S'_{HH(i).}}\right](\nu^\circ)^2 \ \mathrm{var}\ (T|H)$$

as the approximate conditional variance of (11.4) for the special form of the second approach.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Richard S. Levine and Dr. Richard W. Watkins of Educational Testing Serivce, who were responsible for posing the problem to him. The author also benefited from helpful discussions with Professor R. Darrell Bock of the Psychometric Laboratory at the University of North Carolina.

REFERENCES

[1] Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, New York.

[2] Apostol, Tom M. (1957). Mathematical Analysis. Addison-Wesley Publishing Co., Reading, Massachusetts.

[3] Bose, R. C. Mimeographed notes on least squares and analysis of variance. Department of Statistics, University of North Carolina.

[4] Curry, Haskell B. (1944). "The Method of Steepest Descent for Non-linear Minimization Problems". Quarterly of Applied Mathematics, Vol. 2, pp. 258-261.

[5] Gulliksen, Harold (1950). Theory of Mental Tests. John Wiley and Sons, New York.

[6] Householder, Alston S. (1953). Principles of Numerical Analysis. McGraw-Hill Book Co., New York.

[7] Scarborough, James B. (1955). Numerical Mathematical Analysis. The Johns Hopkins Press, Baltimore.

[8] Traub, J. F. (1964). Iterative Methods for the Solution of Equations. Prentice-Hall, Englewood Cliffs, New Jersey.

[9] Tucker, Ledyard R. (1960). Formal Models for a Central Prediction System. Research Bulletin RB-60-14, Educational Testing Service, Princeton, New Jersey.

[10] Zaguskin, V. L. (1961). Handbook of Numerical Methods for the Solution of Algebraic and Transcendental Equations. Pergamon Press, New York.