JOURNAL OF MATHEMATICAL PSYCHOLOGY: 3, 1-18 (1966)

The Axioms and Principal Results of Classical Test Theory

MELVIN R. NOVICK

Educational Testing Service, Princeton, New Jersey

Following an approach due to Guttman the axioms of the classical test theory model are shown to be derivable as constructions from a specified sampling rule and from the assumption that the observed score of an arbitrarily specified or randomly selected person may be considered as an observation of a random variable having finite and positive variance. Without further assumption the reliability of a test is defined. Parallel measurements are then independently defined, and the concept of replication is explicated. The derived axioms of the classical test theory model are then stated in a refined form of Woodbury's stochastic process notation, and the basic results of this model are derived. The assumptions of experimental independence, homogeneity of error distribution, and conditional independence are related to the classical model and to each other. Finally, a brief sketch of some stronger models assuming the independence of error and true scores or the existence of higher-order moments of error distributions or those making specific distributional assumptions is given.

1. INTRODUCTION

The classical theory of mental tests has a long and distinguished history of application to the technology of test construction and test utilization. The most detailed statement of this theory appears in Gulliksen (1950). This theory, however, suffers from some imprecision of statement so that, from time to time, controversies arise that appear to raise embarrassing questions concerning its foundations (Loevinger, 1957; Thorndike, 1964). One purpose of this paper is to show that classical test theory may be placed on a firm theoretical foundation, that its necessary assumptions are very weak and hence generally satisfied. Further it is shown how this model may be modified to provide stronger though less generally applicable models. The function of this partly expository paper is not primarily to derive new results (though some are given) but to explicate the conditions under which old results are valid. In part this presentation parallels and supplements that of Lord (1959b).

2. THE AXIOMS OF THE CLASSICAL TEST THEORY MODEL

By *classical test theory* we shall mean that theory which postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores

and that observed, true and error scores are linearly related. Classical test theory is the simplest case of *weak true score theory*, by which we mean that collection of models that make no specific assumptions concerning the functional form of observed score, true score, or error score distributions.

We consider a denumerably infinite population $\mathscr{P}$ of experimental units (persons) with indexing element a and a denumerably infinite or finite population $\mathscr{I}$ of measurements (tests or items) with indexing element g (the population of measurements possibly containing but a single element). We conceive of the observed score x_{ga}, which is obtained on a specified experimental unit a on a specified measurement g, as one of a number of values that might have been obtained on this experimental unit. We further view this score as being a realization of a real-valued random variable X_{ga}, defined on the set of all possible values that might have been observed over a specified set of conditions, having the cumulative distribution function $F_{g,a}(x_{ga})$, which gives the probability that $X_{ga} \leqslant x_{ga}$. We assume that the variance of the random variable X_{ga} is nonzero and finite. A fixed measurement g is considered to be characterized by the random variables X_{ga}.

Probability statements from $F_{g,a}(x_{ga})$ are not to be interpreted yet as pertaining to relative frequency in repeated trials, as the concept of repeated measurements has not yet been introduced. Under certain further conditions (which we shall indicate later) frequency interpretations are justifiable. At this point we shall be concerned only with the syntactic (mathematical) properties of this distribution and not with its semantic properties. Rather good semantic interpretations of this distribution may be found in Guttman (1945) and in Lazarsfeld (1959).

Denoting a variance by Var we have Var $X_{ga} < \infty$. If $\mathscr{E}$ denotes expectation, then it follows that $\mathscr{E}X_{ga}$ is finite and, for simplicity, may be further denoted as τ_{ga} (an unknown constant) and referred to as the *true score* of experimental unit a on measurement g. Now consider the random variable E_{ga}, taking (unobservable) values e_{ga}, defined by

$$e_{ga} = x_{ga} - \tau_{ga}. \tag{1}$$

Clearly $\mathscr{E}e_{ga} = 0$. We refer to e_{ga} as the error score of experimental unit a on measurement g. Since τ_{ga} is unknown and only x_{ga} is observable e_{ga} is unknown. The error (score) variance, $\sigma^2_{e(ga)}$, and the observed score variance, $\sigma^2_{x(ga)}$, are equal and the correlation between X_{ga} and E_{ga} is unity because of the linear relation (1). The observed, true, and error scores defined above are not those generally considered in test theory. They are, however, those that would be of interest to a theory that dealt with individuals rather than groups (counseling rather than selection).

Now suppose that a measurement g is taken on a randomly selected experimental unit generating the observed score random variable X_{g*} taking values x_{g*}. Let T_{g*} be the random variable (the true score random variable) corresponding to the true score values τ_{g*} that might thus be generated (though not observable) and let E_{g*}

be a random variable (the error score random variable) corresponding to the values e_{g*} thus obtainable. Then clearly

$$e_{g*} = x_{g*} - \tau_{g*}. \tag{2}$$

We say that the random variables X_{g*}, T_{g*}, and E_{g*} are defined in a persons space and denote their variances by $\sigma^2_{x(g*)}$, $\sigma^2_{\tau(g*)}$, and $\sigma^2_{e(g*)}$. We assume the first one of these to be finite and hence, by definition (1), the latter two must also be finite and, to avoid triviality, the latter two are assumed to be strictly greater than zero. Let $\mathscr{E}\tau_{g*} = \pi_g$ (which must be finite). The axioms of classical test theory may then be *derived* from the following theorem.

THEOREM 2.1.

(a) $\mathscr{E}e_{g*} = 0$

(b) $\rho(e_{g*}, \tau_{g*}) = 0$.

If X_{ga} and X_{ha} are independent then E_{ga} and E_{ha} are independent and

(c) $\rho(e_{g*}, e_{h*}) = 0$.

Proof:

(a) $\mathscr{E}e_{g*} = \mathscr{E}[\mathscr{E}e_{ga}] = \mathscr{E}0 = 0$;

(b) Since $\mathscr{E}e_{ga} = 0$ for all (g, a) it will also be zero for a randomly selected a from *any* subset of $\mathscr{P}$. In particular it will be zero for all a such that τ_{ga} is any fixed constant, say τ_{g*}, i.e., $\mathscr{E}(e_{g*} \mid \tau_{g*}) = 0$ for all τ_{g*}. The regression of E_{g*} on T_{g*} is thus linear and the regression line has slope zero, hence *a fortiori* $\rho(e_{g*}, \tau_{g*}) = 0$;

(c) Similarly for fixed measurements g and h, $\mathscr{E}(e_{g*} \mid e_{h*}) = 0$ for all e_{h*} and, hence, $\rho(e_{g*}, e_{h*}) = 0$. The method of proof of sections (b) and (c) fills the gap left by Guttman (1945) and shows quite clearly that E_{g*} and T_{g*} are uncorrelated by *construction* and that the linearity of the model (2) is also a result of construction, and not assumption. The sense in which (c) can be interpreted will be discussed in Section 4.

Our method has provided an analytic decomposition of X_{g*} into the sum of two orthogonal components. This technique is hardly new, mathematically, but to our knowledge it has not found exposition in the psychometric literature with the single exception of Guttman's (1945) perhaps incomplete discussion. Also, this aspect of Guttman's work seems to have been ignored by subsequent writers.

When only one measurement is under discussion and when it is clear that we are referring to a randomly selected a, then, for simplicity, we may omit subscripts and obtain the classical linear test theory model

$$X = T + E, \tag{3}$$

in which X, T, and E are random variables. Classical theory, in its simplest form, deals with X_{g*}; however, extensions of this theory have dealt with X_{*a}, in which random measurements are made on arbitrarily specified experimental units, and X_{**}, in which random measurements are made on randomly selected experimental units. While we have demonstrated that the regression of X on T is linear with unit slope, it is not generally true that the regression of T on X is linear. However, the linear regression line of T on X fitted by the method of least squares will have slope equal to the reliability of the test.

We may introduce the following definitions that complete our foundational structure for the classical model.

Definition I. The reliability of a measurement (test or item) in the population $\mathscr{P}$ is the variance ratio $\sigma^2_{\tau(g*)}/\sigma^2_{x(g*)}$. It is the ratio of between-person variance to total variance and as such depends strongly upon the population $\mathscr{P}$, i.e., the specification of the sampling rule. Each true score defines a reliability, and the various kinds of reliability may be related to various true scores each determined by a specified sampling rule and hence each of the various kinds of reliability can be given explicit definition (Cronbach, Rajaratnam, and Gleser, 1963).

Definition II. Measurements g and g' are parallel (in the classical sense) if, for every $a \in \mathscr{P}$, $\tau_{ga} = \tau_{g'a}$ and $\sigma^2_{e(g'a)} = \sigma^2_{e(g,a)}$. An immediate consequence of this definition is the result $\sigma^2_{e(g*)} = \sigma^2_{e(g'*)}$ for parallel measurements (g, g') and any subpopulation of $\mathscr{P}$. It should be noted that this is a quite strong parallelism requirement in that it requires measurement equivalence for each person and not just average measurement equivalence in some group.

The definition of a true score and the construction of the linear model (3) require only the existence of first moments of the random variable X_{ga}; however, the definitions of reliability and parallelism require the existence of two moments of X_{ga} and X_{g*}. Our assumption that the random variables X_{ga} and X_{g*} have finite variances is a very weak one indeed. Physical (or psychological) measurements are typically bounded, and all positive moments of bounded random variables are finite. Thus, we may say that the classical test theory model has widest applicability.

We now digress from our purely mathematical formulation to relate our model to psychometric application. We assume the (potential) availability of measurements $g', g'', g''', \ldots$, which are parallel measurements, and now consider the random variables X_{ga} and X_{g*} to be associated with the potential values obtained from successive measurements $g', g'', g''', \ldots$. We refer to $g', g'', g''', \ldots$ as *replicate measurements* (with respect to the classical model) and say that X_{ga} is defined in a replications space and X_{g*} is defined in a persons by replications space. The idea here is that "it doesn't matter *a priori* which of the parallel measurements is to be taken at any particular time" since from the point of view of the classical theory (which is concerned only

with means, variances, and covariances) they are equivalent. When this assumption holds, as it often does at least approximately in some applications (i.e., parallel measurements are available), then we have the relation

$$\rho(x_{g'*}, x_{g*}) = \sigma^2_{\tau(g*)}/\sigma^2_{x(g*)} \equiv \sigma^2_\tau/\sigma^2_x \tag{4}$$

say, and the reliability of a test has an equivalent statement as a test-retest correlation coefficient between parallel measurements. Given parallel measurements Eq. 4 provides an operational method of estimating the reliability of a measurement g and statements (a) and (b) of (2.1) are interpretable in the usual test theoretic sense. Indeed, parallel measurements are the central subject of study in classical test theory. The important point to be made, however, is that (as recognized by most psychometricians) the *definition* of reliability does not depend on the concept of equivalence or parallelism, hence, no circularity is involved in these definitions. It is also worthwhile noting that the assumption of experimental independence between measurements on the same person is *not* required, as noted previously by Guttman (1953).

The validity of a measurement g with respect to a (nonparallel) measurement h may be defined as

$$\frac{\sigma_{x(g*)x(h*)}}{\sigma_{x(g*)}\sigma_{x(h*)}} = \rho(x_{g*}, x_{h*}), \tag{5}$$

where $\sigma_{x(g*)x(h*)}$ is the covariance of X_{g*} with X_{h*}. It is clear that this definition is symmetric in g and h. It is well known (and we shall later show, (31)) that

$$\sigma_{x(g*)x(h*)} = \sigma_{\tau(g*)\tau(h*)},$$

where the right-hand member is the covariance of T_{g*} and T_{h*}. Hence an equivalent statement of (5) is

$$\frac{\sigma_{\tau(g*)\tau(h*)}}{\sigma_{x(g*)x(h*)}} = \rho(x_{g*}, x_{h*}). \tag{6}$$

It is apparent that (4) is the limiting case of (6) when g and h are parallel measurements, for then $\sigma_{x(g*)} = \sigma_{x(h*)} = \sigma^2_x$ and (as we shall show, (30), (31)) $\sigma_{\tau(g*)\tau(h*)} = \sigma^2_\tau$.

The weakness of the assumptions of the classical model increases the applicability of this model but restricts its usefulness. The classical test theory model is basically a nonparametric estimation model. We may employ the method of least squares to obtain minimum variance unbiased estimates of true scores and error variances and, by considering the ratios of appropriate estimates of variance, to obtain analog estimates of reliabilities and validities. We may obtain Chebysheff type inequalities (which seem to have been ignored by test theoreticians) and employ classical testing procedures that are distribution free. The more common classical testing procedures, however, require parametric assumptions not provided by this model.

3. A STOCHASTIC PROCESS FORMULATION

The major subject of study of the classical model has been that of the effect of the lengthening of a test on reliability and validity. Recently Woodbury (1963) has given some added detail concerning a formulation of the classical test theory model as the realization of a time dependent stochastic process. By associating the length of a test with a time parameter, t, Woodbury has furnished a model that permits us to provide concise derivations of the major classical test theory results. We shall refine Woodbury's model slightly and derive the major theorems on lengthened tests from it. By adopting Woodbury's model we are able to present these derivations with a considerable degree of preciseness, conciseness, and generality.

We associate the length of a test with a time parameter, t, and view the score of a randomly selected person under specified conditions on a fixed test as the realization of a time dependent stochastic process generating random variables $X(t)$, for $t > 0$. If the test is composed of parallel subtests (items), the domain of t is restricted to the positive integers; if a test consists of performing a task for a time t, then the domain of t is taken as the positive numbers. The basic linear model (derived in essence in the previous section) is

$$x(t) = t\tau + e(t), \tag{7}$$

where $x(t)$, τ, and $e(t)$ are values taken by the observed error and true score random variables $X(t)$, T, and $E(t)$. In this more general formulation the axioms of the classical model may be written as follows:

$$\mathcal{E}[e(t) \mid \tau] = 0, \text{ for all } \tau,\ t > 0; \tag{8}$$

$$\mathcal{E}[\{e(t_2) - e(t_1)\}\{e(t_4) - e(t_3)\}] = 0, \tag{9}$$

when $(t_1, t_2]$ and $(t_3, t_4]$ are disjoint;

$$\mathcal{E}[e(t_2) - e(t_1)]^2 = a^2 \mid t_2 - t_1 \mid,\ 0 < a < \infty, \tag{10}$$

where t_1 and t_2 are values of the argument t.

Assumption (8) corresponds to parts (a) and (b) of the theorem of Section 2. It is a generalization in that it states this property for all $t > 0$. Assumption (9) is a special case of part (c) of the theorem where disjoint t intervals are considered as distinct measurements. More generally we write

$$\mathcal{E}[\{e_g(t_{g_2}) - e_g(t_{g_1})\}\{e_h(t_{h_2}) - e_h(t_{h_1})\}] = 0, \tag{11}$$

where g and h are (possibly) nonparallel measurements. Assumption (10) asserts that measurements defined on time intervals of equal length have equal error variances. This assumption, together with (9), is equivalent to the assumption that the observed

score random variables defined on these intervals are generated by parallel measurements. By considering our expectations to be defined in terms of Stieltjes integrals, we generalize Woodbury's (1963) statement and permit the simultaneous treatment of the discrete and continuous cases.

If $(t_1, t_2] \cap (t_3, t_4]$ is null, let $x_g(t_g) = x(t_2) - x(t_1)$, $x_h(t_h) = x(t_4) - x(t_3)$, $t_g = t_2 - t_1$, and $t_h = t_4 - t_3$. Then $x_g(t_g)/t_g$ may be denoted simply as x_g/t_g. Because of the nature of the model, properties of these difference random variables defined on disjoint t intervals of the same length and pertaining to the same process are identical with each other and with those of such difference random variables generated by distinct but parallel processes. Hence, under these conditions use of the above notation is justified, and such use will be made only under these conditoins. The model we are studying is, perhaps, the simplest case of a weakly stationary (covariance stationary) stochastic process (Parzen, 1962).

Assumptions (9) and (10) may be combined into a single equivalent statement (not given by Woodbury), which has some intuitive value in understanding the nature of the model.

THEOREM 3.1.

$$\mathcal{E}\{[e(t_2) - e(t_1)][e(t_4) - e(t_3)]\} = a^2 t^*, \tag{12}$$

where t^* is the measure of $(t_1, t_2] \cap (t_3, t_4]$, implies and is implied by assumptions (9) *and* (10) *jointly.*

Proof. If $t_4 = t_2$ and $t_2 = t_1$, then (12) reduces to (10); if $(t_3, t_4] \cap (t_1, t_2]$ is the null set, then (12) reduces to (9). The following proof of the converse for one special case is typical of those for other cases. Suppose $t_1 < t_3 < t_2 < t_4$, then by considering disjoint subintervals we have

$$\begin{aligned}\mathcal{E}[\{e(t_2) - e(t_1)\}\{e(t_4) - e(t_3)\}] &= \mathcal{E}[\{e(t_4) - e(t_2)\}\{e(t_2) - e(t_3)\}] \\ &\quad + \mathcal{E}[\{e(t_4) - e(t_2)\}\{e(t_3) - e(t_1)\}] \\ &\quad + \mathcal{E}[\{e(t_2) - e(t_3)\}\{e(t_3) - e(t_1)\}] \\ &\quad + \mathcal{E}[\{e(t_2) - e(t_3)\}^2] \\ &= a^2 t^*.\end{aligned}$$

In the following theorem we develop the basic formulas for expectations and variances and define some of the basic test theory terms.

THEOREM 3.2. *Denoting the distribution functions of E and T by H and G we have:*

$$\mathcal{E}\tau \equiv \pi \text{ (the mean true score);} \tag{13}$$

$$\mathcal{E}[e(t) \mid \tau] = 0, \text{ for all } \tau, \text{ hence } \mathcal{E}[e(t)] = 0; \tag{14}$$

$$\mathscr{E}[x(t)] = \mathscr{E}[t\tau + e(t)] = \mathscr{E}_G\mathscr{E}_H[t\tau + e(t)] + t\mathscr{E}_G\left[\tau + \mathscr{E}_H\frac{e(t)}{t}\right]$$
$$= t\mathscr{E}_G\tau = t\pi; \tag{15}$$

$$\mathscr{E}[x(t_2) - x(t_1)] \equiv \mathscr{E}x_g = (t_2 - t_1)\pi \equiv t_g\pi; \tag{16}$$

$$\mathscr{E}\left[\frac{x(t_2) - x(t_1)}{t_2 - t_1}\right] = \mathscr{E}\frac{x_g(t_g)}{t_g} = \pi; \tag{17}$$

$$\operatorname{Var} e(t) \equiv \sigma^2_{e(t)} = a^2t \quad \text{(the error variance)}; \tag{18}$$

$$\sigma_{e(t)} = at^{\frac{1}{2}} \quad \text{(the error standard deviation)}; \tag{19}$$

$$\operatorname{Var}[e(t)/t] \equiv \sigma^2_{e(t)/t} = a^2t^{-1} \quad \text{(the relative error variance)}; \tag{20}$$

$$\operatorname{Var} x(t) \equiv \sigma^2_{x(t)} = t^2 \operatorname{Var}\tau + \sigma^2_{e(t)} \quad \text{(the observed score variance)}; \tag{21}$$

$$\operatorname{Var}[x(t)/t] \equiv \sigma^2_{x(t)/t} = \sigma^2_\tau + \sigma^2_{e(t)/t} \text{ (the relative observed score variance)}; \tag{22}$$

$$\operatorname{Var}\left[\frac{e(t_2) - e(t_1)}{t_2 - t_1}\right] \equiv \operatorname{Var}\left[\frac{e_g(t_g)}{t_g}\right] \equiv \sigma^2_{e_g/t_g} = a^2 \mid t_2 - t_1 \mid^{-1}; \tag{23}$$

$$\operatorname{Var}\left[\frac{x(t_2) - x(t_1)}{t_2 - t_1}\right] \equiv \operatorname{Var}\frac{x_g}{t_g} \equiv \sigma^2_{x_g/t_g} = \sigma^2_\tau + \sigma^2_{e_g/t_g}. \tag{24}$$

The evaluation of covariances is obtained in the following:

THEOREM 3.3. *Taking* $e(0) = 0$ *with probability one we have*

$$\operatorname{Cov}[e(t_2), e(t_1)] = \mathscr{E}e(t_2)e(t_1) = \sigma^2_{e(t_1)} = a^2t_1\,,\ t_2 \geqslant t_1; \tag{25}$$

$$\operatorname{Cov}\left[\frac{e(t_2)}{t_2}, \frac{e(t_1)}{t_1}\right] = a^2t_2^{-1},\ t_2 \geqslant t_1; \tag{26}$$

$$\operatorname{Cov}\left[\frac{x(t)}{t}, \tau\right] = \mathscr{E}\left[\left\{\tau + \frac{e(t)}{t}\right\}\tau\right] - \mathscr{E}\left[\frac{x(t)}{t}\right]\mathscr{E}(\tau) = \sigma^2_\tau; \tag{27}$$

$$\operatorname{Cov}\left[\frac{x(t)}{t}, \frac{e(t)}{t}\right] = \mathscr{E}\left[\left\{\tau + \frac{e(t)}{t}\right\}\frac{e(t)}{t}\right] = \sigma^2_{e(t)/t}; \tag{28}$$

$$\operatorname{Cov}\left[\frac{x(t_2)}{t_2}, \frac{x(t_1)}{t_1}\right] = \mathscr{E}\left[\left\{\tau + \frac{e(t_2)}{t_2}\right\}\left\{\tau + \frac{e(t_1)}{t_1}\right\}\right]$$
$$- \mathscr{E}\left[\tau + \frac{e(t_2)}{t_2}\right]\mathscr{E}\left[\tau + \frac{e(t_1)}{t_1}\right]$$
$$= \sigma^2_\tau + a^2t_2^{-1},\ t_2 \geqslant t_1; \tag{29}$$

$$\operatorname{Cov}\left[\left(\frac{x(t_2)-x(t_1)}{t_2-t_1}\right),\left(\frac{x(t_4)-x(t_3)}{t_4-t_3}\right)\right]=\sigma_\tau^2+\frac{a^2t^*}{(t_2-t_1)(t_4-t_3)}; \tag{30}$$

$$\operatorname{Cov}\left[\frac{x(t)}{t},\tau_y\right]=\operatorname{Cov}\left[\frac{x(t)}{t},\frac{y(t)}{t}\right]=\operatorname{Cov}[\tau_x,\tau_y]. \tag{31}$$

The proof of (30) is by direct evaluation. The following argument establishes (31):

$$\begin{aligned}\operatorname{Cov}\left[\frac{x(t)}{t},\frac{y(t)}{t}\right]&=\mathscr{E}\left[\left(\tau_x+\frac{e_x(t)}{t}\right)\left(\tau_y+\frac{e_y(t)}{t}\right)\right]\\&\quad-\mathscr{E}\left[\tau_x+\frac{e_x(t)}{t}\right]\mathscr{E}\left[\tau_y+\frac{e_y(t)}{t}\right]\\&=\mathscr{E}\left[\left(\tau_x+\frac{e_x(t)}{t}\right)\tau_y\right]-\mathscr{E}\left[\tau_x+\frac{e_x(t)}{t}\right]\mathscr{E}(\tau_y)\\&=\operatorname{Cov}\left[\frac{x(t)}{t},\tau_y\right]=\mathscr{E}\tau_x\tau_y-\mathscr{E}\tau_x\mathscr{E}\tau_y=\operatorname{Cov}[\tau_x,\tau_y].\end{aligned}$$

The true score of a person has often been defined as the observed score that a person would attain on a very long test. For the present model we may note that the true score, as we have defined it, satisfies this condition since $x(t)/t\to\tau$ in probability as $t\to\infty$. This follows from noting that $\mathscr{E}(x(t)/t)=\tau$ and $\operatorname{Var}x(t)/t=a^2/t\to 0$ as $t\to\infty$ and applying the weak law of large numbers. The mathematical advantage of our choice of definition of true score is obvious and since, from this definition and the assumptions of the model and without any further assumption, we have demonstrated the required asymptotic property we lose nothing (semantically) by our definition.

The basic correlations and their limits are obtained in the following:

THEOREM 3.4. *With $t_2\geqslant t_1$ and $t_4\geqslant t_3$ we have*

$$\rho^2\left(\frac{x(t)}{t},\tau\right)=\frac{\operatorname{Cov}^2\{x(t)/t,\tau\}}{\operatorname{Var}x(t)/t\operatorname{Var}\tau}=\frac{\sigma_\tau^4}{\sigma_\tau^2\sigma_{x(t)/t}^2}=\frac{\sigma_\tau^2}{\sigma_{x(t)/t}^2}; \tag{32}$$

$$\lim_{t\to\infty}\rho^2\left(\frac{x(t)}{t},\tau\right)=\lim_{t\to\infty}\frac{\sigma_\tau^2}{\sigma_\tau^2+(a^2/t)}=1; \tag{33}$$

$$\lim_{t\to 0}\rho^2\left(\frac{x(t)}{t},\tau\right)=0; \tag{34}$$

$$\rho^2\left(\frac{x(t)}{t},\frac{e(t)}{t}\right)=\frac{\operatorname{Cov}^2\{x(t)/t,e(t)/t\}}{\operatorname{Var}x(t)/t\operatorname{Var}e(t)/t}=\frac{\sigma_{e(t)/t}^2}{\sigma_{x(t)/t}^2}; \tag{35}$$

$$\lim_{t\to\infty} \rho^2\left(\frac{x(t)}{t}, \frac{e(t)}{t}\right) = \lim_{t\to\infty} \frac{1}{(t\sigma_\tau^2/a^2) + 1} = 0; \tag{36}$$

$$\lim_{t\to 0} \rho^2\left(\frac{x(t)}{t}, \frac{e(t)}{t}\right) = 1; \tag{37}$$

$$\rho^2\left[\frac{x(t_2) - x(t_1)}{t_2 - t_1}, \frac{x(t_4) - x(t_3)}{t_4 - t_3}\right] = \frac{\{\sigma_\tau^2 + [a^2 t^*/(t_2 - t_1)(t_4 - t_3)]\}^2}{\{\sigma_\tau^2 + [a^2/(t_2 - t_1)]\}\{\sigma_\tau^2 + [a^2/(t_4 - t_3)]\}}. \tag{38}$$

When intervals are nonoverlapping and $t_4 - t_3 = t_2 - t_1$ we have

$$\rho\left(\frac{x_g}{t}, \frac{x_h}{t}\right) = \sigma_\tau^2/\sigma_{x(t)/t}^2, \qquad \text{or} \qquad \sigma_\tau^2 = \sigma_{x(t)/t}^2 \rho(x_g/t, x_h/t); \tag{39}$$

from (32) and (39) we have

$$\rho^2(x(t)/t, \tau) = \rho(x_g/t, x_h/t); \tag{40}$$

from (32), (35), and (39) we have

$$\rho^2(x(t)/t, e(t)/t) = [1 - \rho(x_g/t, x_h/t)]$$

and

$$\sigma_{e(t)/t}^2 = \sigma_{x(t)/t}[1 - \rho(x_g/t, x_h/t)]. \tag{41}$$

In the remainder of this section we consider measurements of unit length, i.e., $t = 1$, and denote $x(t)/t$ by x. A measurement of length $kt = k$ will be denoted by $x(k)$. Following this notation the results needed to relate the whole and a part of a test are given in the following:

THEOREM 3.5.

$$\mathscr{E}\frac{x(k)}{k} = \mathscr{E}\frac{[k\tau + e(k)]}{k} = \pi, k = 1, 2, 3, \ldots; \tag{42}$$

$$\begin{aligned}
\operatorname{Var}\frac{x(k)}{k} &= \operatorname{Var}[k^{-1}([x(1) - x(0)] + [x(2) - x(1)] + \cdots + [x(k) - x(k-1)])] \\
&= \operatorname{Var}\left[k^{-1}\sum_{g=1}^{k} x_g\right] \\
&= k^{-2}\left[\sum_{g=1}^{k} \operatorname{Var} x_g + \sum_{h=1}^{k}\sum_{g=1}^{k} \operatorname{Cov}(x_g, x_h)\right] \\
&= k^{-2}[k\sigma_x^2 + k(k-1)\sigma_x^2\rho(x_g, x_h)] \\
&= \frac{\sigma_x^2}{k}[1 + (k-1)\rho(x_g, x_h)];
\end{aligned} \tag{43}$$

$$\operatorname{Cov}\left[\frac{x(k)}{k}, x\right] = \mathscr{E}\left(\frac{x(k)}{k}\right)(x) - \mathscr{E}\left(\frac{x(k)}{k}\right)\mathscr{E}(x)$$

$$= k^{-1}\mathscr{E}\left[x_g^2 + x_g \sum_{h=2}^{k} x_h\right] - \mu^2$$

$$= k^{-1}\sigma_x^2[1 + (k-1)\rho(x_g, x_h)]; \tag{44}$$

the square of the part-whole correlation is then

$$\rho^2(x(k), x) = k^{-1}[1 + (k-1)\rho(x_g, x_h)]. \tag{45}$$

The proofs of the results of Theorem 3.5 have assumed that k was a positive integer. If k is not, but if, in lowest terms, $k = a/b$, where a and b are integers, then t/b may be taken as unit length and the above and all succeeding derivations are valid. If k is irrational, the results will still hold by a limiting argument.

The Spearman-Brown formula for the reliability of a lengthened test is given in the following:

THEOREM 3.6.

$$\rho^2\left(\frac{x_g(k)}{k}, \frac{x_h(k)}{k}\right) = \frac{k^2\rho^2(x_g, x_h)}{[1 + (k-1)\rho(x_g, x_h)]^2}. \tag{46}$$

Proof:

$$\operatorname{Cov}\left[\frac{x_g(k)}{k}, \frac{x_h(l)}{l}\right] = (kl)^{-1}\mathscr{E}\left[\left(\sum_{g=1}^{k} x_g\right)\left(\sum_{h=1}^{l} x_h\right)\right] - \mu^2$$

$$= \sigma_x^2\rho(x_g, x_h),$$

which does not depend on k or l. This, together with (43), establishes (46).

More generally suppose that x_g and y_h are generated by distinct processes and suppose that the square of the correlation between them is $\rho^2(x_g, y_h)$, then

THEOREM 3.7.

$$\rho^2\left(\frac{x_g(k)}{k}, \frac{y_h(l)}{l}\right) = \frac{kl\rho^2(x_g, y_h)}{[1 + (k-1)\rho(x_g, x_{g'})][1 + (l-1)\rho(y_h, y_{h'})]}, \tag{47}$$

where (g, g') and (h, h') are pairs of parallel measurements.

The general result (47) can then be employed to obtain the standard results on the validity of a test as a function of test lengths which are contained in the following:

THEOREM 3.8. *The effect of multiplying the length of the predictor by k may be obtained by taking* $l = 1$ *in* (47). *Then*

$$\rho^2\left(\frac{x_g(k)}{k}, y_h\right) = \frac{k\rho^2(x_g, y_h)}{[1 + (k-1)\rho(x_g, x_{g'})]}. \tag{48}$$

The square of the correlation between predictor and criterion when the predictor is of infinite length (and hence perfectly reliable) is

$$\lim_{k\to\infty} \rho^2\left(\frac{x_g(k)}{k}, y_h\right) = \lim_{k\to\infty} \frac{\rho^2(x_g, y_h)}{[(1/k) + (1 - (1/k))\rho(x_g, y_h)]} = \frac{\rho^2(x_g, y_h)}{\rho(x_g, x_{g'})}. \tag{49}$$

The square of the correlation between predictor and criterion when both predictor and criterion are of infinite length (and hence perfectly reliable) is

$$\lim_{\substack{k\to\infty \\ l\to\infty}} \rho^2\left(\frac{x_g(k)}{k}, \frac{y_h(l)}{l}\right) = \frac{\rho^2(x_g, y_h)}{\rho(x_g, x_{g'})\rho(y_h, y_{h'})}. \tag{50}$$

The square of the correlation between predictor and criterion when the criterion is of infinite length (and hence perfectly reliable) is

$$\lim_{l\to\infty} \rho^2(x_g, y_h) = \frac{\rho^2(x_g, y_h)}{\rho(y_h, y_{h'})}. \tag{51}$$

Formulas (49), (50), and (51) are often referred to as attenuation formulas.

4. EQUIVALENT MEASUREMENTS AND EXPERIMENTAL INDEPENDENCE

The classical test theory model deals entirely with means, variances, and covariances. When comparing measurements within the context of this theory other properties of these measurements are not considered. Parallel measurements were defined as those which, within the context of the classical model, were interchangeable (i.e., equivalent) in the sense that "it did not matter which of the measurements were taken." We refer to the concept of parallelism as an equivalence relationship. Each test theory model requires its own particular equivalence relationship. For example, a somewhat stronger model might require that equivalent measurements have the same true scores and the same second and third order moments of the error distributions. Yet stronger models would require equivalent measurements to have the same true scores and identically distributed error distributions.

When the random variable X_{ga} is considered to be defined in a replications space of equivalent (parallel) measurements, we may adopt a completely deterministic

point of view (which has previously been suggested by Cronbach (1947) and others) concerning the (so-called) error random variable E_{ga}. We conceive of the error random variable as a composite of a multitude of factors that are not controlled in the experiment. Error variance is simply that variance which is not attributable to variances among conditions of the experiment. From this point of view, for a fixed person and a population of experimental conditions, the error random variable may be considered to be a nondegenerate random variable in the replications space, while the true score may be considered to be a degenerate random variable (a constant) for fixed a. Also, given equivalent measurements we may refer to the true score random variable as being defined in a persons space or equivalently in a persons by replications space.

While results (a) and (b) of Theorem 2.1 require only the assumption of finite variances, a second assumption is of importance in establishing part (c) of the theorem. This is the assumption of experimental independence. Measurements g and g' are said to be experimentally independent if for every $a \in \mathscr{P}$ the joint distribution function $F_{gg',a}(x_{ga}, x_{g'a})$ factors into the product $F_{g,a}(x_{ga})F_{g',a}(x_{g'a})$ of the marginal distribution functions. Since we are dealing here with a fixed individual a, it is readily seen that the assumption of experimental independence is implied by the assumptions that the error random variables E_{ga} and $E_{g'a}$ are statistically independent and that the true scores τ_{ga} and $\tau_{g'a}$ are unaffected by previous measurement, i.e., that neither the person nor the physical environment is affected by the first measurement. Part (c) of Theorem 2.1 can be interpretable in the usual test theoretic sense when measurements g and h are assumed to be experimentally independent. Several approximately experimentally independent observations are usually available in mental testing work.

The following example may be of some value in evaluating the assumption of experimental independence. Suppose that we administer a test consisting of a number of items to a subject who happens to have a very bad headache and suppose that this headache causes the subject to do poorly on the test. Now with respect to a vector of true scores corresponding to the various items and considering a hypothetical random sample of headache–nonheadache conditions for the subject, what we have obtained is a consistent pattern of negative error scores caused by the fixed headache condition. Clearly there is a correlation among these error scores. Error scores between measurements on a fixed subject will be uncorrelated only when the sampling rule guarantees experimental independence of the measurements for each subject. Cronbach (1947) and others have discussed the conditions under which this assumption is reasonable in practice.

Given replications in terms of identical true score and identically distributed error random variables for all a, and given experimental independence among trials, the distribution function $F_{g,a}(x_{ga})$ is then descriptive of the long run relative frequency of the various possible values x_{ga} obtained from such trials. While long runs of such replications are not obtainable generally this interpretation is valid theoretically and useful conceptually.

5. HOMOGENEITY OF ERRORS AND CONDITIONAL INDEPENDENCE

Further strengthening of the classical model has been accomplished by introducing the assumption of conditional independence. We shall describe and state this assumption in a form consistent with the developments in the previous sections of this paper.

Consider any finite set $\mathscr{I}'$ of specified measurements from $\mathscr{I}$ and a randomly selected person. The associated random variables X_{g*}, X_{h*}, etc. are defined in a persons (by replications) space. Consider then that subpopulation $\mathscr{P}'$ of persons with specified fixed true τ_g, $\tau_{h'}$, etc., i.e., those persons whose expected observed scores with respect to the random variables X_{ga}, $x_{g'a}$, etc. for an arbitrary $a \in \mathscr{P}'$ are all equal to the specified quantities τ_g, $\tau_{h'}$, etc. Let

$$\boldsymbol{\theta} = (\tau_g, \tau_h, \ldots)$$

be the vector containing a finite number of elements τ.

Within $\mathscr{P}'$, or any subpopulation (finite or infinite) of $\mathscr{P}'$, a strong form of the assumption of local (conditional) independence states that the joint distribution of X_{g*}, $X_{g'*}$, etc. factors into the product of the marginal distribution functions. Notationally this may be expressed by the following equation.

$$F_{gh\ldots,*}(x_{g*}, x_{h*}, \ldots \mid \boldsymbol{\theta}) = F_{g,*}(x_{g*} \mid \boldsymbol{\theta}) F_{h,*}(x_{h*} \mid \boldsymbol{\theta}) \ldots . \tag{52}$$

Concerning local independence Anderson (1959) wrote:

Apart from any mathematical reason for such an assumption there are psychological or substantive reasons. The proposition is that the latent quantities [true scores] are the only important factors and that once these are determined behavior is random (in the sense of statistical independence). In another terminology, the set of individuals with specified latent characteristics are (sic) "homogeneous."

Anderson does not pursue this discussion further. An excellent heuristic discussion of this may be found in Lazarsfeld (1959). This discussion also suggests that the assumption of experimental independence is part of the assumption of local independence.

In the context of the model development in this paper the homogeneity condition may be formally stated by the equation:

$$F_{g,a}(x_{ga} \mid \boldsymbol{\theta}) = F_{g,a'}(x_{ga'} \mid \boldsymbol{\theta}), \tag{53}$$

where a and a' are arbitrary elements from $\mathscr{P}'$. It would seem appropriate to refer to this as the assumption of homogeneous errors. This assumption states that all persons with the same true scores have the same error distributions in the replications space.

Now, in an obvious notation,

$$F_{gg',*} \mid \boldsymbol{\theta} = \mathscr{E}_a F_{gg',a} \mid \boldsymbol{\theta}.$$

By experimental independence

$$F_{gg',a} = F_{g,a} \cdot F_{g'a},$$

hence

$$F_{gg',a} \mid \boldsymbol{\theta} = F_{g,a} \mid \boldsymbol{\theta} \cdot F_{g',a} \mid \boldsymbol{\theta}.$$

Thus

$$F_{gg',*} \mid \boldsymbol{\theta} = \mathscr{E}_a[F_{g,a} \mid \boldsymbol{\theta} \cdot F_{g',a} \mid \boldsymbol{\theta}].$$

By the assumption of homogeneous errors

$$F_{g,a} \mid \boldsymbol{\theta} \equiv F_{g,a'} \mid \boldsymbol{\theta}, F_{g',a} \mid \boldsymbol{\theta} \equiv F_{g',a'} \mid \boldsymbol{\theta},$$

hence the expectation, $\mathscr{E}_a$, is taken over a constant and hence

$$F_{gg',*} \mid \boldsymbol{\theta} = F_{g,a} \mid \boldsymbol{\theta} \cdot F_{g',a} \mid \boldsymbol{\theta}$$

or

$$F_{gg',*} \mid \boldsymbol{\theta} = F_{g,*} \mid \boldsymbol{\theta} \cdot F_{g',*} \mid \boldsymbol{\theta}.$$

Thus the assumption of conditional independence is implied by the joint assumptions of experimental independence and homogeneity of errors.

The converse statement is also true for, given conditional independence,

$$\mathscr{E}_{a|\theta} F_{gg',a} \equiv F_{gg',*} \mid \theta = F_{g,*} \mid \theta \cdot F_{g',*} \mid \theta = \mathscr{E}_{a|\theta} F_{g,a} \cdot F_{g',a}$$

and this statement must hold for all $(x_g, x_{g'})$ and for *any subset of* $\mathscr{P}'$. In particular it must hold for all pairs of $a \in \mathscr{P}'$. If this expectation equation is to hold for arbitrary subsets (including pairs) then the values taken as a function of $(x_g, x_{g'})$ must be constant for all a. Hence

$$F_{gg',a} \mid \theta = F_{gg',a'} \mid \theta,$$

and hence

$$F_{g \cdot a} \mid \theta = F_{g \cdot a'} \mid \theta.$$

Thus the assumption of conditional independence implies the assumption of homogeneous errors. Furthermore

$$F_{gg',a} \mid \boldsymbol{\theta} = F_{g,a} \mid \theta \cdot F_{g',a} \mid \theta \text{ or } F_{gg',a} = F_{g,a} \cdot F_{g',a} \text{ in } \mathscr{P}',$$

and hence the assumption of conditional independence implies the assumption of experimental independence. Thus the assumption of conditional independence is equivalent to the assumptions of experimental independence and homogeneity of errors taken jointly. It is also worth noting that the assumption of homogeneous errors is equivalent to the assumption that the distribution of X_{ga} belongs to a parametric class of distributions and that the true score τ_{ga} may serve as an indexing element for that class. This assumption would seem to be a reasonable one if the situation

is such that a person's errors are determined by factors associated with the measuring technique but not with the individual. However, if measurements are such that the responses of different persons with the same true score have differing variability under identical external conditions, then the assumption of conditional independence is a tenuous one.

Latent structure theory and factor analytic theory may be considered as natural extensions of true score theory. Each of these theories is concerned with the reduction in the dimensionality of the vector θ of true scores. Assuming that these true scores are functionally related and hence expressible in terms of some lesser number of "latent variables," latent structure theory proposes a number of models that assume various metrics for these latent variables and various kinds of relationships among them. The latent linear model of latent structure theory is just the Spearman single factor model.

6. FURTHER MODELS

Throughout the test theory literature the assumption that true and error scores are uncorrelated has been, whenever necessary, strengthened to assert that the scores were independently distributed. An assumption equivalent to that of independence is the assumption that the conditional random variable E, given a fixed value $T = \tau$, has the same distribution H for all τ, i.e., $H(E \mid \tau) = H(E)$. From our basic equation $X = T + E$ we obtain $X \mid \tau = \tau + E \mid \tau$ and $X \mid \tau - \tau = E \mid \tau$, where $X \mid \tau$ and $E \mid \tau$ are the conditional observed and error random variables. Since under the assumption that E and T are independent the distribution of $E \mid \tau$ does not depend on τ, it follows that the distribution of $X \mid \tau - \tau$ does not depend on τ, or, equivalently, we say that τ is a location parameter for X. This terminology is justified in that the distribution of $X \mid \tau$ for the various values τ is the same except for location, which is determined by τ. Thus, since the converse argument is valid we see that the assumption of independent true and error scores is equivalent to the assumption that τ is a location parameter for X.

Under the assumption of independently distributed errors both the strong law of large numbers and the central limit theorem are applicable for large values of t, and hence, asymptotically, $[(x_{ga}/t_{ga}) - \tau_{ga}]/\sigma_{x(ga)}$ has the standard normal distribution and x_{ga}/t_{ga} converges to τ_{ga} with probability one. This result needs to be distinguished from similar results describing the asymptotic behavior of the sample mean of observations of n persons. In this case, if errors between persons are independent, then the behavior of this average over persons will be governed by both the central limit theory and the strong law of large numbers. This latter error independence assumption is surely the more tenable and indeed as demonstrated in the previous section is equivalent to our statement of the assumption of local independence.

In Section 3 we obtained the basic result Var $e(t) = a^2t$, which followed immediately from assumption (10). This variance homogeneity property of the classical model may be strengthened in the following way. Denote the distribution function of $E_g(t_g) = E(t_2) - E(t_1)$ by $H_g(e_g)$. We assume then that $H_g(e_g) = \Psi([t_2 - t_1]^{-1/2}e_g(t_g))$, where the function Ψ does not depend on (t_1, t_2). It is easily seen that this homogeneity property implies the previously stated one but is indeed a much stronger assumption. When the stronger property holds we say that $(t_2 - t_1)^{1/2}$ is a scale parameter for H. The idea here is that the error difference distributions are identical except for scale.

If the location and scale assumptions hold jointly, i.e., if the distribution of $[x(t) - \tau]/\sqrt{t}$ does not depend on τ or t, then we say that the model for X is a regression model. Regression models have a very interesting predictive frequency property, which has been presented in considerable generality by Hall and Novick (1964). Of the more commonly used distributions the normal and negative exponential distributions are regression models while the binomial and Poisson are not.

All models considered to this point have involved assumptions concerning only the first two moments of the error distribution. Some investigations, however, have employed distribution free models assuming the existence and centering attention on higher order moments, see, for example, Lord (1959a). Since, under mild regularity conditions, the moments of a distribution determine the distribution, it is possible to study true and error score distributions by estimating their moments. As with means and variance, the assumption of the existence of higher order moments is justifiable on the basis of the boundedness of physical measurement. Some very useful results relating higher order moments of observed and latent variables under the assumption of local independence have been given by Anderson (1959).

A final step in the strengthening of the classical model is to assume a particular functional form for the error distribution, or for the conditional observed score distribution for fixed true score. A survey of some important uses to which the normal and binomial models can be put has been given by Lord (1959b). Generally the primary justification for the assumption of specified distribution forms has been empirical, not theoretical. (An exception is Rasch, 1960.) These models have been accepted in part because the assumptions from which they may be derived seem reasonable for the process under study but more generally they are accepted because "they fit the data." A systematic investigation of the theoretical bases for the assumptions of various strong models is beyond the scope of our present undertaking.

ACKNOWLEDGMENT

I wish to thank Frederic M. Lord for his encouragement and guidance of this work and for his incisive criticism of early drafts of this material. This criticism, together with that of the referee have, in my judgment, contributed substantially to the clarity of presentation of the material.

REFERENCES

ANDERSON, T. W. Some scaling models and estimation procedure in the latent class model. *Probability and statistics. The Harold Cramer Volume.* New York: Wiley, 1959. Pp. 9-38.

CRONBACH, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, **12**, 17-30.

CRONBACH, L. J., RAJARATNAM, N., AND GLESER, GOLDINE C. Theory of generalizability: a liberalization of reliability theory. *British Journal of statistical Psychology*, 1963, **16**, 137-163.

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255-283.

GUTTMAN, L. Reliability formulas that do not assume experimental independence. *Psychometrika*, 1953, **18**, 225-239.

HALL, W. J., AND NOVICK, M. R. A note on classical and Bayesian prediction intervals for location, scale, and regression models. Mimeograph Series, Institute of Statistics, University of North Carolina, 1964.

LAZARSFELD, P. Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science*, Vol. 3. New York: McGraw-Hill, 1959.

LOEVINGER, JANE. *Objective tests as instruments of psychological theory.* Missoula, Montana: *Psychological Reports*, 1957.

LORD, F. M. Statistical inferences about true scores. *Psychometrika*, 1959, **24**, 1-18. (a)

LORD, F. M. An approach to mental test theory. *Psychometrika*, 1959, **24**, 283-303. (b)

PARZEN, E. *Stochastic processes.* San Francisco: Holden-Day, 1962.

RASCH, G. *Studies in mathematical psychology. I. Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson and Lydiche, 1960.

THORNDIKE, R. L. Reliability. *Proceedings of the 1963 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1964. Pp. 23-32.

WOODBURY, M. A. The stochastic model of mental testing theory and an application. *Psychometrika*, 1963, **28**. 391-394.

RECEIVED: December 21, 1964