UTILITY FUNCTIONS FOR TEST PERFORMANCE

James S. Dyer
William Farrell
Paul Bradley

Center for the Study of Evaluation
University of California, Los Angeles

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 3-7, 1972.

ABSTRACT

This paper discusses a utility function estimation procedure developed to provide curriculum planning information to elementary school principals. Both theoretical and empirical studies were performed to evaluate the procedure. The results of the use of this approach to obtain data from a national sample of principals are presented.

1. INTRODUCTION

This paper discusses a utility function estimation procedure developed to provide curriculum planning information to an elementary school principal. In a previous paper, Dyer [7] described how this information may be used to rationalize the selection of curriculum goal areas (e.g., reading comprehension) for relative emphasis. Here we shall confine our discussion to the procedure used to obtain utility function estimates from a non-random national sample of elementary school principals.

The utility function we wish to estimate is defined on test scores of a particular group of students.[1] We assume that their performance may be characterized by an N-vector of scores $s = (s_1,\ldots,s_N)$. Each component of this vector measures group performance in one of the N goal areas composing a school's curriculum. The most common presentation of the result from a standardized test is in terms of a percentile score from [0,100] determined relative to a norming population.[2] Thus, we define $S_n = [0,100]$, $n = 1,\ldots, N$, as the set of all possible percentile scores in goal area n, and we require $s_n \in S_n$. Let $S = (S_1\times\ldots\times S_N)$ be the set of all possible N-vectors of percentile scores $s = (s_1,\ldots,s_N)$. We assume that the elementary school principal's preference relation $\prec$ weakly orders score vectors on S.[3] We wish to estimate a real valued function U on S satisfying $U(s^1) < U(s^2)$ if and only if $s^1 \prec s^2$ for N-vectors $s^1, s^2 \in S$.

---

[1] For a complete discussion of this formulation, see Dyer [7]. A less technical discussion is given by Amor and Dyer [1].

[2] See Hoepfner, et al., [13] for a critical review of the existing test instruments for measuring student performance in curriculum goal areas.

[3] A weak order $\prec$ on S is assymmetric and negatively transitive.

The outline of this paper is as follows. In section two, we greatly simplify our problem by assuming that the decision maker's utility function is additively separable in the criteria. This allows us to consider each criterion independently, and to estimate a single dimensional utility function defined on each. The results can then be aggregated by additive weighting. Although we have replaced one large problem by N smaller ones, we are now in a position to choose from several relatively simple utility function estimation procedures. We then explain why we selected the "direct ordered metric" method for estimating these single-criterion utility functions.

The third section deals with the investigation of several questions regarding the adequacy of the direct ordered metric procedure. These questions were raised by Davidson, Suppes, and Siegel [4], by DeGroot [6], and by others. Both theoretical and empirical results are given which indicate that "good" estimates can be obtained from this procedure. In the fourth section, the results of the use of this method with a national sample of principals are presented. Finally, the fifth section provides the conclusions.

2. SELECTION OF PROCEDURE

2.1 Additively Separable Utility Functions

The task of estimating a utility function defined on multiple criteria (in our case, N different test scores) is particularly difficult without prior knowledge regarding the function's form. An important simplification results if we assume that the desired utility function is additively separable in the criteria. This assumption implies the existence of N real valued functions $u_1,\ldots,u_N$ on $S_1,\ldots,S_N$ respectively such that $U(s) = \sum_{n=1}^{N} u_n(s_n)$ for $s = (s_1,\ldots,s_N)$.

Debreu [5] provides three conditions required for this additivity assumption when decisions are made over criteria whose values are known with certainty. Two of these conditions, topological requirements on the S_n, may be demonstrated easily for $S_n = [0,100]$, $n = 1,\ldots,N$. The third condition is the utility independence of the criteria. Loosely speaking, this means that the decision maker's preference relation defined on any one of the criteria must be independent of the values of the other criteria.

When the values of the criteria are uncertain, one further condition is needed to justify this additive separability assumption. In such cases, Fishburn [8] has shown that this form of utility function requires the decision maker to be indifferent between all lotteries having equal marginal probabilities for the criteria values. The relationship between the Debreu-independence and the Fishburn-marginality conditions is examined in Sec. 2.3, and the implications of this relationship for our procedure are described.

These conditions do indicate that a decision maker is likely to exhibit an additively separable utility function when good performance in one goal area cannot be perceived as compensating for poor performance in any other goal area. The 41 curriculum goal areas we consider (see Hoepfner, et al., [11] for descriptions) have been identified so that each represents a distinct skill, cognitive ability, or affective trait. Hence, we have defined these goal areas in a manner intended to minimize any competitive or complimentary effects on the decision maker's utility function. The process followed in the development of this list of goals was similar to the procedure detailed by Pardee, et al., [16], and resulted in a hierarchial goal structure. If student performance is measured in these 41 areas, we do not expect the error resulting from the additive separability assumption to

be significant. Thus, for the remainder of this paper, we assume that the required conditions for additive separability are satisfied.

A convenient form of an additive separable function is obtained by defining

$$u_n(s_n) \equiv w_n f_n(s_n) \qquad n = 1,\ldots,N \tag{1}$$

where w_n is a positive constant for each n, and f_n is a real valued function. Since each f_n is defined on the domain $S_n = [0,100]$, we may arbitrarily restrict its range to [0,1] by requiring $f_n(0) = 0$ and $f_n(100) = 1$ for $n = 1,\ldots,N$. Note that a score of 0 is the absolute worst obtainable in a goal area, and that 100 represents the absolute best. Therefore, the w_n's may be interpreted as reflecting the "relative importance" of performance in each goal (see Miller [15]).

Several methods for estimating these scale factors (the w_n's) have been identified (see Fishburn [9]). The Center for the Study of Evaluation (CSE) at UCLA has developed a technique which allows elementary school principals to determine their own values of the scale factors for the 41 curriculum goal areas [11]. The instructions include suggestions for utilizing information obtained from teachers and parents. This technique was used successfully by elementary school personnel in a nationwide field test [12].

The estimation of the N functions f_n is a more difficult matter. Miller [15] and Pardee, et al., [16] provide instructions for the estimation of "worth" functions by the decision maker. However, practical considerations suggest that, if our procedure is to be considered useful

by the typical elementary school principal, this rather arduous task must be performed for him. The remainder of this paper will be concerned with that task.

2.2 The Estimation Procedure

A preliminary group of elementary school principals were interviewed to determine the most appropriate utility estimation procedure. Fishburn [9] lists eight different approaches for the estimation of f_n. Four of these introduce the notion of a lottery or "standard gamble" to elicit the desired information. The principals in the preliminary sample responded unfavorably to questions involving probabilities of achieving specific test results. They indicated, for example, that a "50 percent chance of obtaining a test score in word comprehension of 40 percentile" was not a meaningful concept. Keeney [14] has noted that operational difficulties may occur when decision makers are required to express a preference for a lottery instead of directly in terms of the criteria. It is possible that this difficulty could have been overcome by a trained interviewer. However, since the use of an interviewer was not practical in our situation, and since we wished to avoid elaborate instructions, no approach involving lotteries received further consideration.

Of the remaining four techniques, the process of "ranking" would have yielded no useful information. We believed it reasonable to assume that an elementary school principal would find a higher score at least as preferable as a lower score. This implies that f_n will be monotonically increasing in s_n for each n, and provides a ranking of the test score values. The direct rating of attribute values (on an arbitrary scale) and the direct identification of a midpoint[4] were also presented to the principals. While

[4] An example would be "Select the score $\bar{s}_n$ which you consider twice (or one-half) as 'good' as s_n^*."

they reported less difficulty in relating to the questions, they indicated a lack of confidence in their ability to provide the responses.

The approach generally favored by the sample group of principals is termed "direct ordered metric" by Fishburn [9]. This procedure requires a ranking of utility differences among discrete criterion values.[5] We selected seven points (15, 30, 40, 50, 60, 70, 85) from the domain of each f_n. By estimating the value of f_n at each of these seven points, we obtain a piecewise linear estimate of the desired function.

The selection of these particular seven points was determined from an empirical study guided by several heuristic rules. Since we did not wish to make a priori assumptions (other than monotonicity) about the form of the utility functions, we insisted that the points be distributed symmetrically about the midpoint. However, we were concerned that principals might consider the national norm (by definition, 50th percentile) to be an "aspiration level," and indicate much stronger preferences for scores above this point than below. Therefore, we were relatively more concerned with obtaining accuracy towards the center of the test score range.

As additional points are added, the amount of information required to obtain accurate estimates increases. After trying other possibilities, we chose the seven points to determine two 15 percentile intervals on either side of four 10 percentile intervals. As we shall see, several empirical studies indicated that responses to fifty questions relating the intervals determined by pairs of these seven points were sufficient to provide "good" estimates

---

[5] An axiomatic basis for a utility theory based on the concept of preference differences is presented by Suppes and Winet [19].

of known curves. This result was considered an acceptable compromise between the desire to reduce information requirements and the need for accurate estimates. The fifty questions are of the following type:

> Your students have just taken a nationally standardized test in creativity. The test has two parts, A and B, and they represent two aspects of the subject that are equally important to you. Test results are in percentile scores for school norms. Your school averages were
>
> Part A 50 percentile
>
> Part B 70 percentile
>
> Which increase would be worth more to you--
>
> Part A from 50 to 60 percentile
>
> or
>
> Part B from 70 to 85 percentile?

The response to each question of this type yields an inequality relationship. For example, the response "A" to the above question would indicate that

$$f_n(60)-f_n(50) \geq f_n(85)-f_n(70) \tag{2}$$

where n represents the goal area. The goal area in this case is creativity. Similar questionnaires were used for other goal areas.

If the decision maker responds to the questionnaire consistently,[6] the resulting fifty inequalities together with monotonicity constraints ($f_n(x_1) \geq f_n(x_2)$ whenever $x_1 \geq x_2$ and $x_1, x_2 \in \{0, 15, 30, 40, 50, 60, 70, 85, 100\}$) and two equality constraints ($f_n(0)=0$, $f_n(100)=1$) form a convex set. Each point in this set is of the form $(0, f_n(15),\ldots,f_n(85), 1)$ and represents a piecewise linear approximation to the desired utility function. No other conditions regarding the shape of the function (e.g., concavity) were imposed.

---

[6] No single response can be considered inconsistent. However, a response indicating that $f_n(85)-f_n(70) \geq f_n(50)-f_n(30)$ is inconsistent with the response $f_n(50)-f_n(40) \geq f_n(85)-f_n(70)$ if $f_n(40)-f_n(30)>0$.

Since the data were acquired by questionnaire, it was anticipated that inconsistent responses would occasionally be obtained. In such a case, the convex set defined by the fifty-seven constraints generated from the responses to the questionnaire and the monotonicity requirements may be empty. Therefore, the non-negative variable θ was introduced into all constraints, so that (2), for example, would be written

$$f_n(60)-f_n(50) + f_n(70)-f_n(85) + \theta \geq 0. \tag{3}$$

If θ is minimized via linear programming, a value of $\theta = 0$ will be obtained if the constraints written according to (2) are consistent, while a solution that is "closest" to being consistent in some sense is produced if the minimum θ is greater than zero.

One interpretation of a positive value for θ is that it represents a "threshold of preference"; that is, if the utility difference between scores x_1 and x_2 differs by more than θ from the utility difference between x_3 and x_4, then the respondent will always indicate $f(x_1) - f(x_2) \geq f(x_3) - f(x_4)$. Otherwise, there is some positive probability that the inequality will be reversed [4]. As we shall see, some empirical tests were performed to investigate the behavior of the estimation procedure under conditions allowing this interpretation.

2.3 Assumptions of the Questionnaire

Fishburn [8] has noted that (2) is equivalent to

$$.5f_n(60) + .5f_n(70) \geq .5f_n(85) + .5f_n(50) \tag{4}$$

which would arise if questions regarding a 50-50 lottery were used. However, he criticizes approaches based on the concept of preference differences on the grounds that they violate the requisite hypothesis that S_n be a set of mutually exclusive outcomes ([10], p. 81). To overcome this difficulty, we

have "specified" performance in each goal area n by performance in two subgoals. We then ask the decision maker to express his preference over differences between percentile scores in each of these two subgoal areas.

Let us generalize this concept as follows. Suppose we have some criterion C and some measure of performance on this criterion. This measure of performance will be some real number $x \in X \subseteq R^1$. We now introduce "sub-criteria" $C_1,\ldots,C_M$, and with each sub-criterion C_m we associate a performance measure $y_m \in Y_m \subseteq R^1$. For a particular decision maker, we say the collection of M sub-criteria $C_1,\ldots,C_M$ specifies the criterion C if he believes that knowledge of performance in each of these sub-criteria is sufficient to determine performance in C. That is, $\{C_1,\ldots,C_M\}$ specifies C if the decision maker believes there is a real valued (perhaps unknown) function g such that $g(y_1,\ldots,y_M) = x$ for all $x \in X$ and all $(y_1,\ldots,y_M) \in Y_1\times\ldots\times Y_M$. Such a collection $\{C_1,\ldots,C_M\}$ is called a specification of C. Finally, let v denote a utility function on X. Clearly, if $\{C_1,\ldots,C_M\}$ specifies C, there exists some utility function $\tilde{v}$ on $Y_1\times\ldots\times Y_M$ such that $\tilde{v}(y_1,\ldots,y_M) = v(g(y_1,\ldots,y_M))$.

This concept of a specification is a formalization of Raiffa's notion of a partition of goals [17]. Loosely speaking, these conditions ensure that we may determine the decision maker's utility function for a criterion either directly (if possible) or via a vector of performance measures in a set of sub-criteria which specify this criterion.

With this concept, it is a straightforward matter to demonstrate the following: Given the criterion C whose performance measure is $x \in X \subseteq R^1$ and the M sub-criteria $C_1,\ldots,C_M$ whose performance measures are $y_m \in Y_m \subseteq R^1$ for $m = 1,\ldots,M$ respectively, if

C1. $X = Y_m$ for $m = 1,\ldots,M$;

C2. the M sub-criteria C_m, $m = 1,\ldots,M$, are a specification of C;

C3. $\tilde{v}(y_1,\ldots,y_M) = \tilde{v}_1(y_1) + \cdots + \tilde{v}_M(y_M)$ (additive separability of $\tilde{v}$);

C4. $\tilde{v}_i = \tilde{v}_j$ for all $i, j \in \{1,\ldots,M\}$; and

C5. $g(\bar{x},\ldots,\bar{x}) = \bar{x}$ for all $\bar{x} \in X$;

then $v(x) = M\tilde{v}_m(x)$ for all $x \in X$ and all $m \in \{1,\ldots,M\}$. Thus each $\tilde{v}_m$ is identical to v except for the positive constant M.

Let us now consider how these comments apply to our questionnaire. We have indicated in our scenario that parts A and B represent a specification of the particular goal area (C2). The student performance in the goal area and in each of these two sub-goal areas is measured in terms of percentiles (C1). By indicating that the two parts of the test instrument represent equally important (but undefined) aspects of the test, we expect that the use of the additive separable form $f_n = f_{n1} + f_{n2}$ will be a reasonable approximation (C3), that $f_{n1} = f_{n2}$ (C4), and that $g(\bar{x},\bar{x}) = \bar{x} \in S_n$(C5). Thus, we expect $f_n = 2f_{n1} = 2f_{n2}$, and the general form of (2) is equivalent to

$$(2') \qquad f_{n1}(\bar{s}_{n1}) - f_{n1}(\bar{\bar{s}}_{n1}) \geq f_{n2}(\bar{s}_{n2}) - f_{n2}(\bar{\bar{s}}_{n2})$$

for any scores $\bar{s}_{n1},\bar{\bar{s}}_{n1},\bar{s}_{n2},\bar{\bar{s}}_{n2} \in S_n$.

The results from our study will be used in a situation involving decision making under uncertainty. We have avoided utility estimation procedures which

explicitly involve uncertainty in the form of lotteries. Raiffa [17] distinguishes those functions which are a guide to decisions only in non-probabilistic choice situations by terming them value functions. We now wish to consider whether our estimates will be appropriate for decisions made under uncertainty, or whether they are valid for use only as value functions.

Referring again to our previous definitions, let $\tilde{u}$ denote a probabilistic utility function on $Y_1 \times \ldots \times Y_M$, and $\tilde{v}$ denote a value function on $Y_1 \times \ldots \times Y_M$. It is known that the Fishburn-marginality conditions for decision making under uncertainty imply the Debreu-independence conditions for decision making with certain outcomes (see Raiffa [17], p. 56). Thus, if there exists $\tilde{u} = \tilde{u}_1 + \cdots + \tilde{u}_M$, then $\tilde{u} = \tilde{v}$ up to a positive linear transformation. However, it is also clear that if $\tilde{v} = \tilde{v}_1 + \cdots + \tilde{v}_M$ is determined under conditions of certainty, and if the Fishburn-marginality conditions are satisfied, $\tilde{v} = \tilde{u}$ up to a positive linear transformation. Formally, we have the following result.

Proposition. If the Fishburn-marginality conditions are met, and if the M value functions $\tilde{v}_1, \ldots, \tilde{v}_M$ are determined such that $\tilde{v}(y_1, \ldots, y_M) = \tilde{v}_1(y_1) + \cdots + \tilde{v}_M(y_M)$, then $\tilde{v}$ is also a probabilistic utility function. The proof is straightforward.

The implication of this result is important. It allows us to avoid explicitly introducing uncertainty into a utility estimation procedure, and yet still be justified in using the resulting estimate as a guide to decision making under uncertainty. We must, of course, empirically verify or assume that the Fishburn-marginality conditions are satisfied.

3. ADEQUACY OF THE ESTIMATES

Several questions have been raised in the literature regarding the accuracy of estimates obtained by this procedure. Since, to our knowledge, these questions had never received further consideration, the ability of this technique to provide "good" estimates of utility functions was studied.

One of the issues considered was the lack of a unique solution to the set of constraints. Hence, the ability of the questionnaire to reproduce both deterministic and stochastic versions of several test curves was investigated. The latter experiment provided insight into the size and meaning of a nonzero value for θ.

3.1 Approximation to Centroid

The procedure discussed in this paper has been criticized by DeGroot [6] on the grounds that, in general, there is no unique solution to the linear programming problem which minimizes θ. The minimum value of θ may be compatible with a convex polyhedron of solutions for the values of f.[7] In such a case, Davidson, Suppes, and Siegel [4] suggest that the centroid of the polyhedron would provide the preferred solution. However, due to "computational limitations," they selected the first feasible solution obtained by the linear programming routine.

The choice of the centroid of the convex polyhedron is intuitively appealing but computationally arduous. This solution was approximated by using the values from two linear programming problems as follows. Let $x_i, i = 1,\ldots,7,$ correspond to the seven percentile scores (exclusive of 0 and 1) related in the questionnaire. Subject to the constraints derived

[7] In this section, the notation has been simplified by deleting the subscript n.

from the questionnaire, let $f^*(x_i)$, i=1,...,7 and θ^* be an optimal solution to the problem

$$\text{maximize} \quad \sum_{i=1}^{7} f(x_i) - M\theta$$

and let $f_*(x_i)$, i=1,...,7 and θ_* be an optimal solution to

$$\text{minimize} \quad \sum_{i=1}^{7} f(x_i) + M\theta$$

where M is a large number. The piecewise linear estimate of the utility function is obtained from the nine points $\hat{f}(x_i)$ by requiring $\hat{f}(0)=0$, $\hat{f}(100)=1$, and by letting $\hat{f}(x_i)=\frac{1}{2}(f^*(x_i) + f_*(x_i))$, i=1,...,7; clearly the values of $\hat{f}(x_i)$ are a feasible solution to the constraints generated from the questionnaire (for minimum $\theta=0$). Further, it is a straightforward matter to demonstrate that if the constraints are consistent (i.e., there exists at least one feasible solution with $\theta=0$), then $\theta^*(\theta_*) \leq 7/M$. Otherwise, if $\theta^*(\theta_*) > 7/M$, the constraints are inconsistent.

It would be desirable to compare estimates obtained from this procedure with those obtained from using the centroid. Unfortunately, the determination of the centroid requires the identification of all of the extreme points of the convex polyhedron. The number of possible basic solutions provides an upper bound on the number of extreme points. In our case, this number is well over 500,000,000.[8]

---

[8]There is no available procedure for determining a priori the number of extreme points in a convex polyhedron. However, the number of possible basic solutions is given by $\binom{64}{7}$ = 621, 216, 192. The 64 represents the number of constraints (including the non-negativities) and the 7 is the number of variables.

As an alternative to the centroid, the results of the estimation procedure above were compared with the estimate determined from the centroid of a 7-cell found by maximizing and minimizing in each of the coordinate directions. This is the minimum 7-cell containing our convex set. Formally, we determine this 7-cell by solving the fourteen linear programming problems maximize (minimize) $f(x_i)$ for $i = 1,\ldots,7$ subject to the constraints generated by the questionnaire and the monotonicity assumption.

The non-uniqueness of the solution is of concern primarily in the case where $\theta^*=\theta_*=0$. Otherwise, a unique result is likely. Therefore, eleven questionnaires with "consistent" responses ($\theta^*=\theta_*=0$) were used to investigate the difference in the approximations determined by the two methods. These questionnaires had been filled out by principals participating in our national sample.

In eight of the eleven cases, the values of the $\hat{f}(x_i)$ coincided exactly with the values determined from the centroid of the 7-cell. The other three cases are shown in Figure 1, where the solid line is the estimate from the centroid of the 7-cell and the dashed line is from the $\hat{f}(x_i)$. It is also important to note that these three curves represent estimates of the actual utility functions of principals participating in our study.

Thus, the estimates obtained from the fourteen extreme point solutions did not differ significantly from those obtained from the approach using only two extreme points. Although this result is not equivalent to a comparison with the actual centroid, it does suggest that the convex polyhedron formed by the procedure does not exhibit a pathological structure which obviates the intent of using the values of the $\hat{f}(x_i)$. However, we still must investigate whether the use of our approximation procedure actually does provide an acceptable estimate to a utility function.

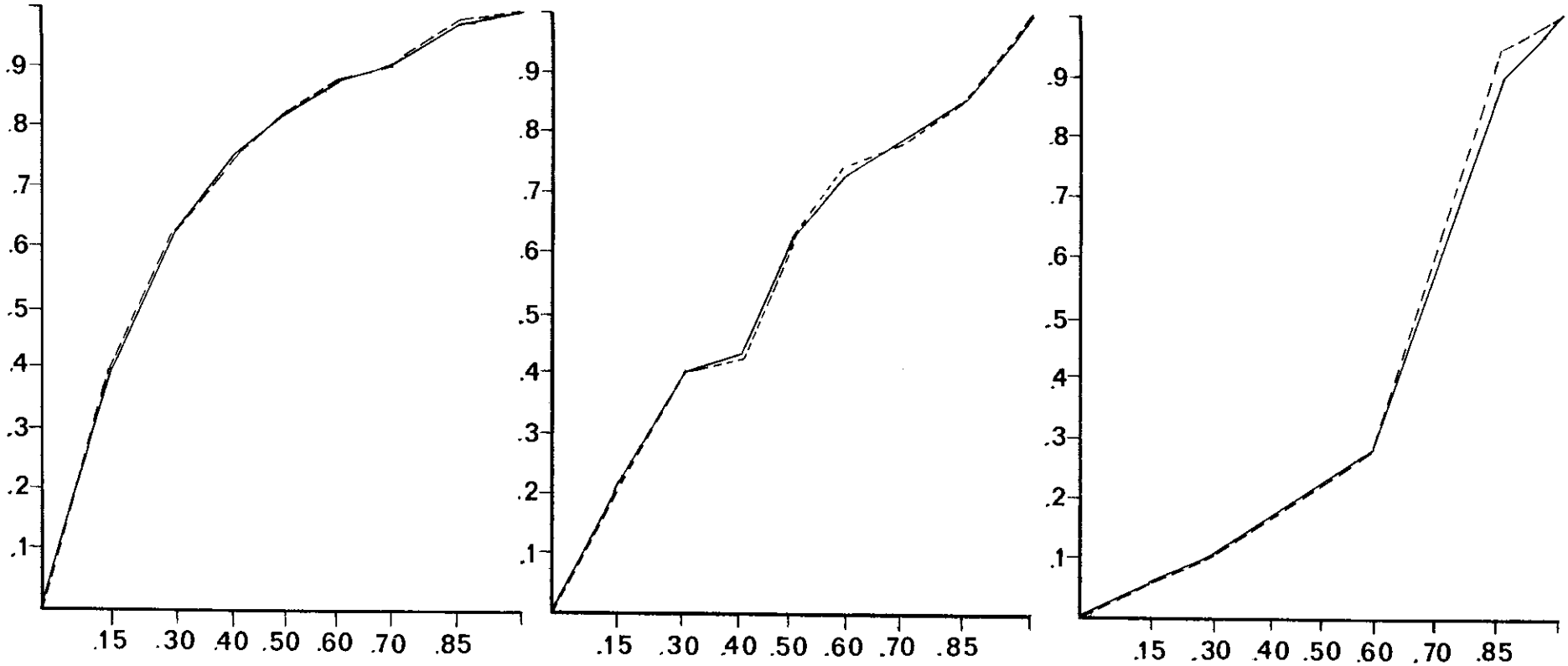


Figure 1. A Comparison of Estimates

3.2 Estimates of Deterministic Functions

The most desirable test of our procedure would compare the estimate obtained from the values of the $\hat{f}(x_i)$ with the decision maker's actual utility function. Since the latter is unavailable, an alternative test was developed. A questionnaire of the type used in the field study was answered according to the responses implied by each of twenty-four test curves. These curves, each defined over the closed interval [0,100], were restricted only by monotonicity, by f(0)=0, and by f(100)=1. Included were a straight line, several convex curves, several concave curves, several s-shaped curves, and ten piecewise linear curves whose coordinates (f(15),...,f(85)) were generated randomly on the computer subject only to the above restrictions.

In a critical analysis of the work of Davidson, Suppes, and Siegel, DeGroot [6] suggested that the estimates provided by their procedure were no better in predicting subject responses than an actuarial curve (a straight line joining (0,0) and (100,1)). Therefore, the actuarial curve was selected as a standard for comparison.

The responses to the questionnaire for each of the twenty-four test curves were used to generate piecewise linear estimates of the original functions. Comparisons between the estimates and the corresponding original curves were made at the seven points 15, 30, 40, 50, 60, 70, 85. The mean square difference, total absolute difference, and maximum absolute difference were computed. These same measures were calculated for the difference between the test curves and the actuarial curve. Finally, the areas between the test curves and the corresponding estimates, and between the test curves and the actuarial curve were approximated. Examples of four of the test curves and their estimates are shown in Dyer [7].

The actuarial curve proved to be a better approximation for only two of the original curves. Naturally, when the test curve was linear, the actuarial curve fit perfectly. However, in this case the mean square difference for the estimate from our procedure was 0.00043 (recall that the range of the curve is 1.0), and the maximum difference was only 0.035. For the other curve better approximated by the actuarial function, neither provided a particularly good fit. The estimate from our procedure was better than the actuarial curve on every test for twenty of the twenty-four curves. For the other two, the estimate was superior for every test except the "maximum absolute difference" which is a single point estimate.

Using the nonparametric sign test for matched pairs, results were sufficient to reject at the .01 level the null hypothesis that the fit provided by the actuarial curve was as good as our estimate. They strongly suggest that <u>if</u> the decision maker does have preferences for test scores which may be described by a well-defined, deterministic utility function, and <u>if</u> he has correctly responded to the questionnaire in terms of these preferences, then our procedure will provide a "good" estimate to this function.

3.3 Estimates of Stochastic Functions

The previous discussion has been based on the assumption that subjects respond consistently in terms of a well-defined, deterministic utility function. In practice, it is known that persons do not invariably make the same choice when faced with the same options. In our study, only 11 of the 60 responding subjects returned questionnaires for which $\theta^*, \theta_* = 0$.

Two basic alternatives are to consider all inconsistent responses as errors, or to define preference and indifference in terms of probabilities

of choice (see Davidson and Marshak [3]). The latter case may be simulated by introducing a stochastic error term associated with an underlying utility function. Using this approach, we have investigated the effects of inconsistent responses on our estimation procedure.

Six curves were selected for study. Included were a concave, a linear, two convex, and two s-shaped curves. For each curve, the error term was assumed to be normally distributed (with $\mu=0$) and truncated at $\pm 3\sigma$. Three different values of σ, 0.01, 0.05, and 0.1, were used. The approach can best be explained in terms of an example. Suppose the question deals with the following choice:

Part A from 70 to 85 percentile

or

Part B from 40 to 60 percentile

For the concave test curve (see Figure 2), $f(60)=.86$, $f(40)=.74$, $f(85)=.96$, and $f(70)=.91$. In the deterministic case explained in Sec. 3.2, the response to this question would be "Part B." Now, suppose we select the values of f from normal distrbutions with $\sigma=.05$ about the true values of the curve. Thus, we may obtain $\bar{f}(60)=.824$, $\bar{f}(40)=.725$, $\bar{f}(85)=1.000$, and $\bar{f}(70)=.893$ where $\bar{f}$ denotes the estimate of f obtained from sampling. In this case, the response to the question would be "Part A," which is clearly incorrect for the underlying function.

This procedure was used for each question in the questionnaire. An important result of this experiment was the generation of numerous estimates with $\hat{\theta}>7/M$. These non-zero values were in the same range as those found in many of the responses from actual subjects. This finding is significant for three reasons. First, this information does not contradict the assumption that the inconsistent responses of many of the subject principals result from

Figure 2a. Linear Curve (σ=.01)

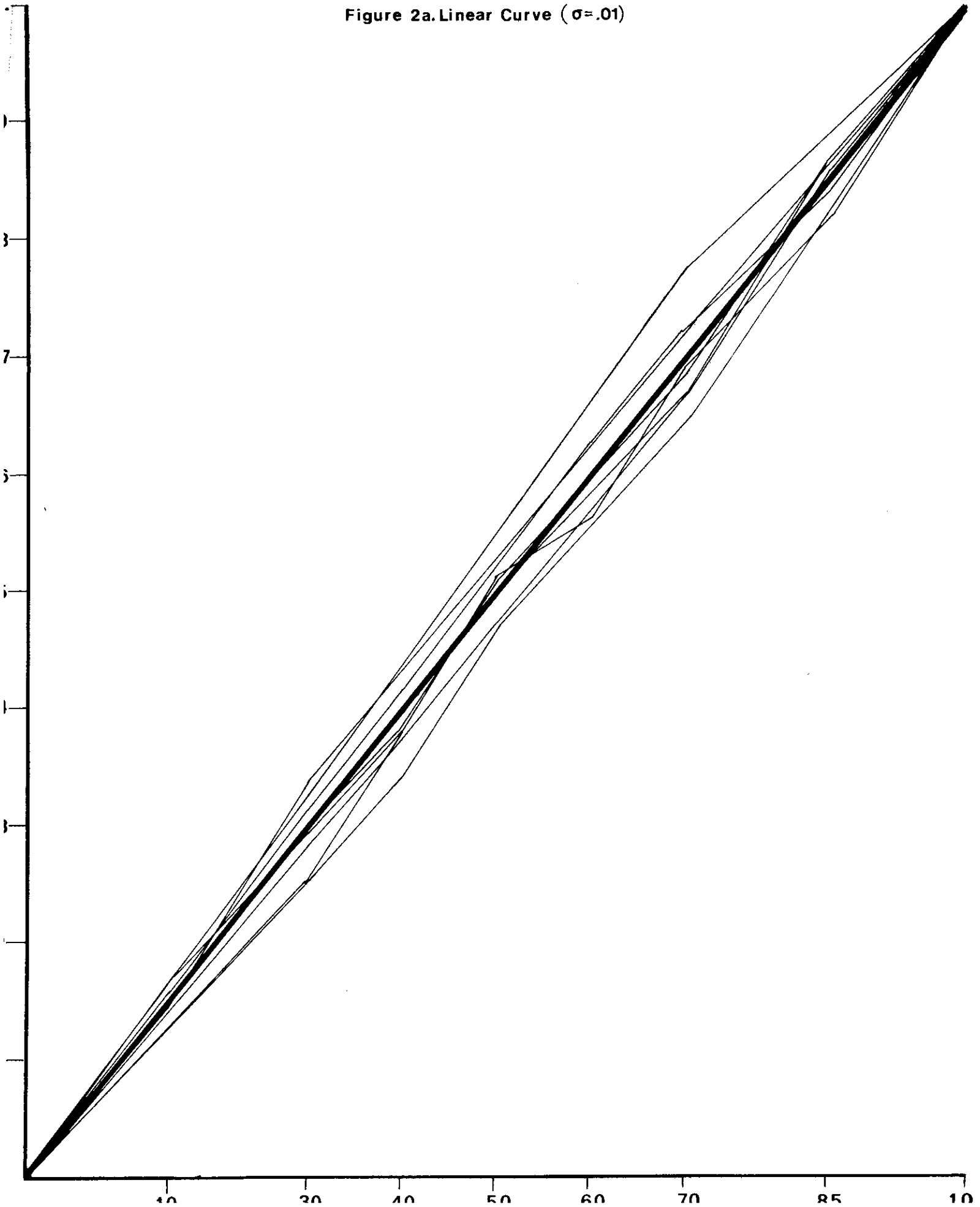


Figure 2b. Concave Curve (σ=.05)

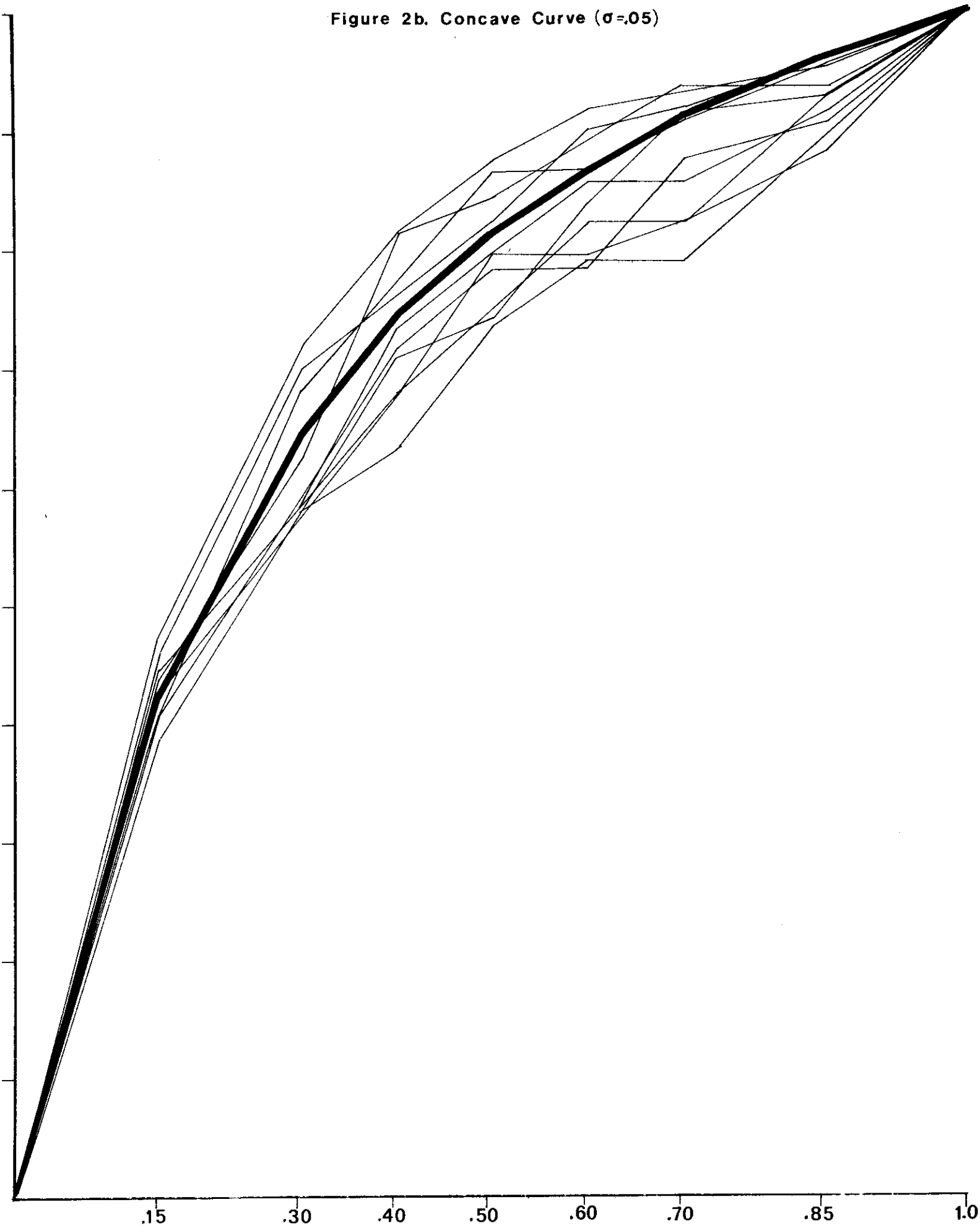


Figure 2c. Convex Curve (σ=.05)

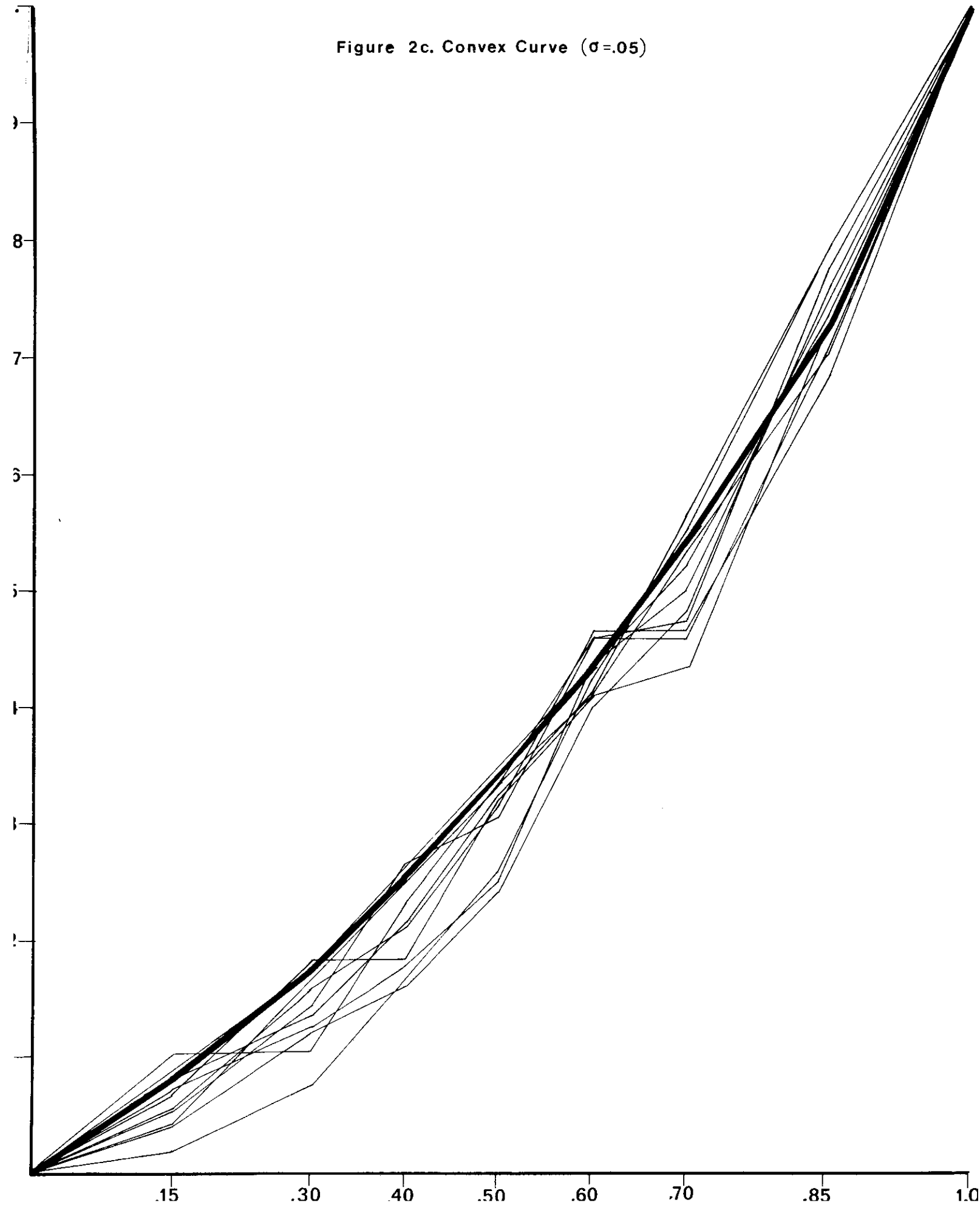
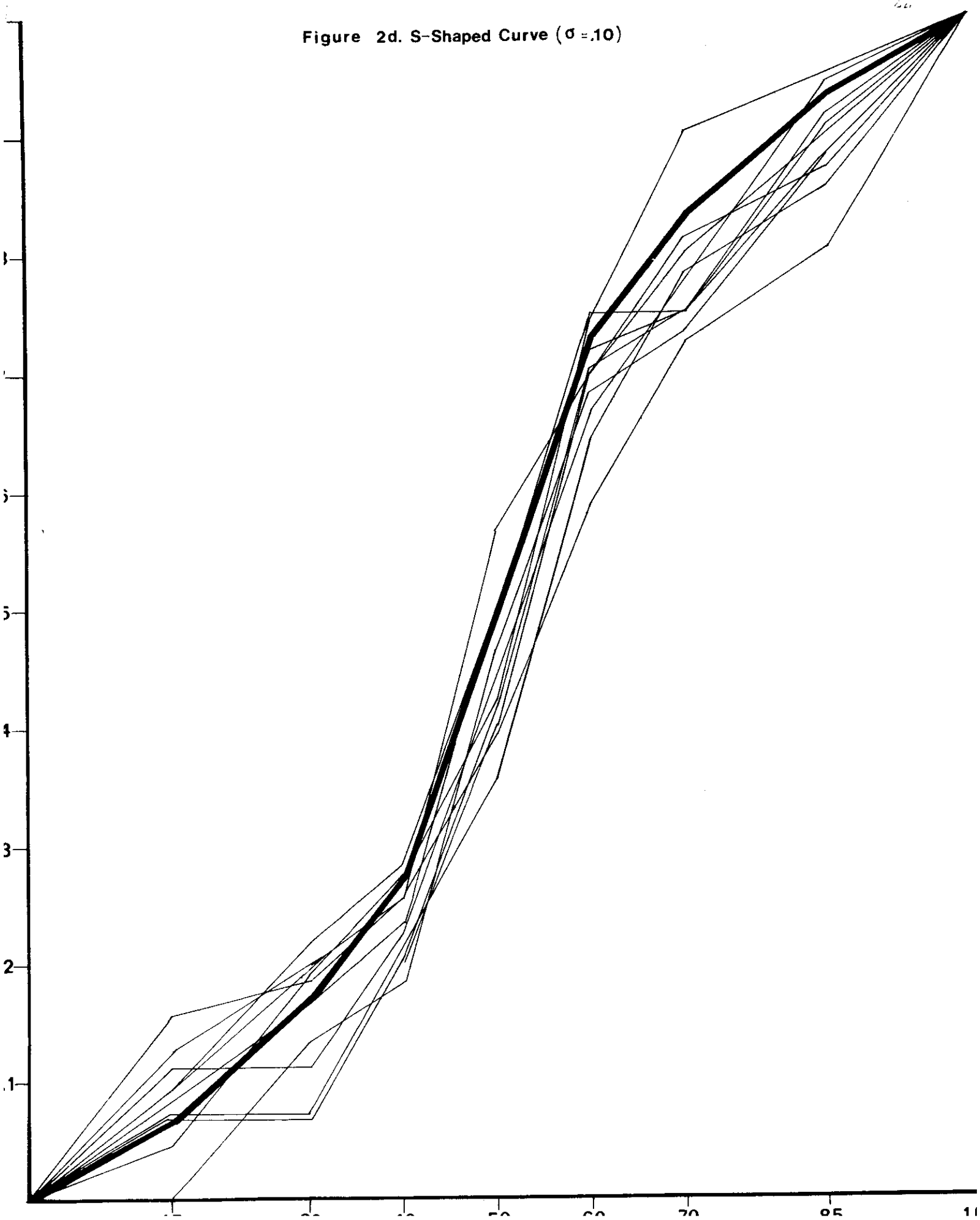


Figure 2d. S-Shaped Curve ($\sigma = .10$)



a "fuzzy" or stochastic utility function. Second, this gives us some indication of the sense in which the "minimum Θ" solution is "close" to a consistent solution. Third, and most important, the estimates provided by many of the solutions with $\Theta > 7M$ (indicating that inconsistent responses were recorded) were still good approximations to the underlying curve.

A sample of the results is shown in Figure 2. The value of σ is also indicated for each curve.

4. EVALUATION OF QUESTIONNAIRE RESPONSES

This section describes the use of our estimation procedure to obtain data from a national sample of elementary school principals. The questionnaire was sent to 72 principals, and of the 62 that were returned, 60 were usable. Each questionnaire specified one of twelve educational goal areas to which the principal should relate his judgements. The goal areas selected for this survey are listed in Table 1.

As a basis for aggregating the responses, we assumed that the principal's judgements would be influenced primarily by the particular goal area under consideration and by his aspiration level (see Siegel [18]) for student performance. As reported by Coleman, et al., [2], certain demographic variables are highly correlated with typical levels of student performance on nationally standardized tests. Thus, for a given goal area n, we hypothesized that the utility function f_n would be similar for principals of schools with similar demographic characteristics.

Demographic information about the schools of principals that participated in the study enabled us to classify each school into one of three school types. School type 1 had demographic characteristics that would predict low student achievement, school type 2 had demographic characteristics that would predict average student achievement, and school type 3 had demographic characteristics that would predict high student achievement. The number of questionnaires that were usable for this cross classification of goals and school types is given in

Table 1

Goal Names and Cell Frequencies

	School Type		
	1	2	3
Scientific Processes	3	4	1
Social Temperament	0	3	1
Sociology	1	5	0
History & Civics	0	2	3
Scientific Knowledge	1	6	2
Physical Education	3	3	2
Attitudes	1	2	1
Arithmetic Operations	1	1	1
Arithmetic Concepts	1	1	2
Reading Comprehension	2	2	1
Language Construction	0	0	1
Creativity	0	2	1

Table 1. The principals who returned questionnaires represent a cross section of the United States, with about half being from California and with the remainder being from several other states.

In order to determine if there were any statistically significant differences among the utility functions, a two way multivariate analysis of variance was performed.[9] The two ways were educational goals (12 levels) and school types (3 levels), and the dependent variables were the utility curve values at seven points between 0 and 100 (15, 30, 40, 50, 60, 70, 85). Because there are six empty cells in the design, and because there are unequal cell frequencies, it was necessary to perform a least squares, non-orthogonal, multivariate analysis of variance.

The only significant effect appearing in the results was the main effect of goals. A plot of the mean utility curves for the 12 goal areas is shown in Figure 3. Nine of the twelve curves are very similar in shape. The convex curve (for the goal area Language Construction) is the most deviant of the 12 curves, but it is based only on a sample of one (see Table 1). This one case was eliminated from the sample, and the analysis was performed again. The results of this analysis on the 11 x 3 design are given in Table 2, and it is easily seen that there are no significant effects. Thus, these results indicate that there are no statistically significant differences among the curves attributable to the effects of either the 11 goal areas or the 3 school types.

While there was some disappointment that the expected differences among utility curves on the basis of goal area and school type did not prove to be statistically significant, the explanation may lie with the relatively small sample size. The small sample size diminishes the power of the analysis to detect "real" differences. It is still possible that utility functions quite different from the mean utility curve could be obtained from sampling schools that are known to be extremely low or extremely high in student achievement. Recall that we classified

---

[9]The program used was MULTIVARIANCE, written by Jeremy D. Finn, SUNY, Buffalo.

Figure 3 Curves By Goal Area

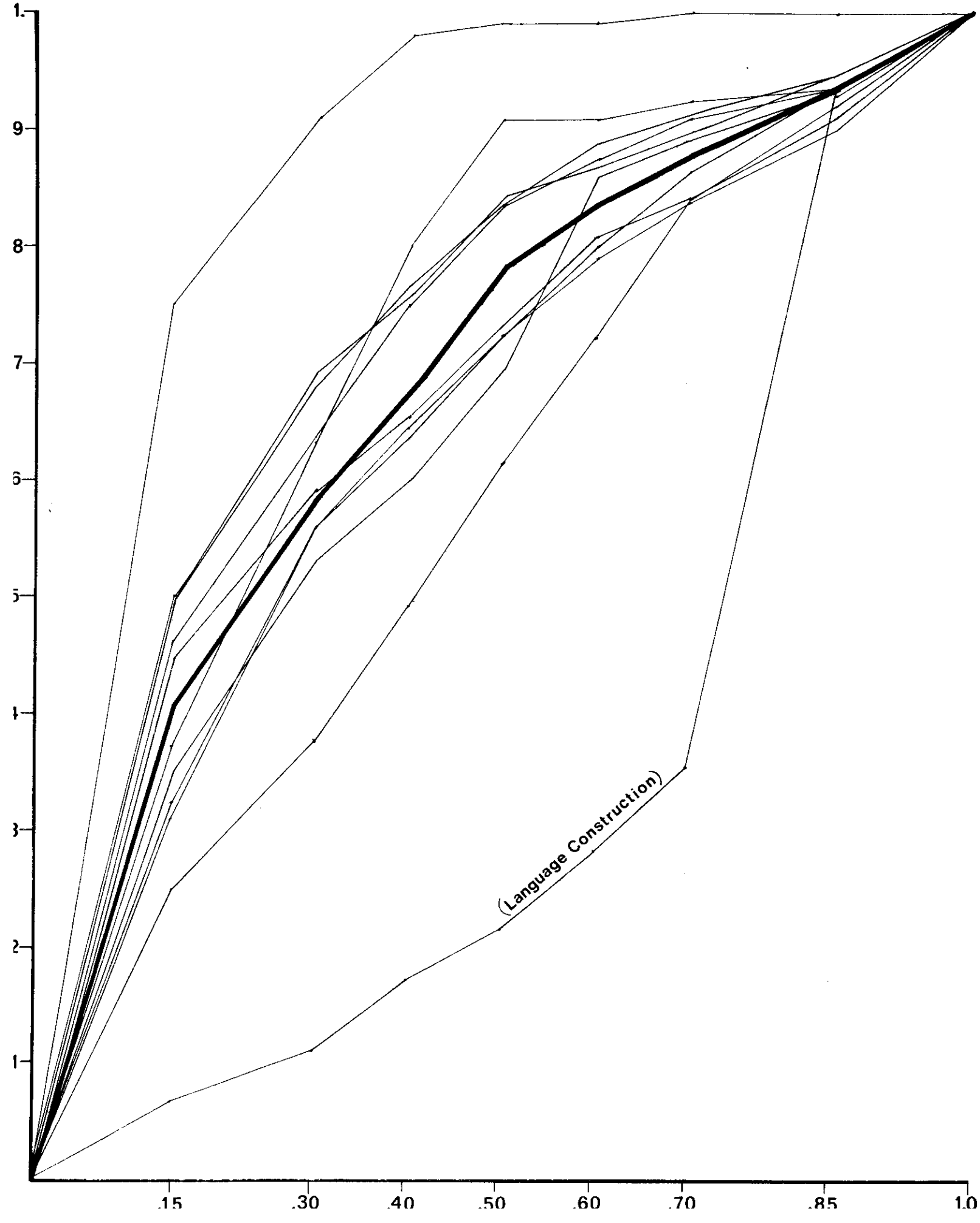


Table 2

Multivariate Analysis of Variance: 11 Goals

Source	Unbiased F	df Hypothesis	df Error	P
Goals	1.0792	70	146.76	<0.3459
School Types	1.2849	14	48	<0.2515
Goals x School Types	0.8573	112	164.82	<0.8083

schools according to the predicted level of student achievement.

The mean utility curve for all principals (the heavy line in Figure 3) can be closely approximated by three linear segments over the following three ranges of test performance: 0-15, 15-50, and 50-100. The difference in utility values for 0 and 15 is about equal to the difference in utility values for 15 and 50, which is about 0.4. This result indicates that elementary school principals, no matter what type of school they administer, and no matter what the educational goal area is, associate the greatest value with improving student performance from the worst possible score to the 15th percentile. Of nearly equal value, but involving a larger difference in student achievement, is improving student performance from the 15th to the 50th percentile. It is not surprising that the slope of the utility function changes abruptly at the 50th percentile, since the 50th percentile is the "national average" and becomes a "target" or aspiration level. A principal would probably experience less criticism if his school's performance is at least average than if his school's performance is below average. It is interesting to note that this mean utility curve exhibits the smooth concavity consistent with the "law of diminishing marginal utility" empirically verified in numerous studies in an economic context.

5. CONCLUSIONS

The information regarding elementary school principals' utility functions has been incorporated into a simple procedure for selecting educational goal areas for relative emphasis (see Dyer [7] and Hoepfner, et al., [11]). Using data based on an even smaller preliminary sample, the procedure has been nationally field tested. Evaluations by principals were generally favorable, although some complained about the degree of "over-sophistication." Additional efforts are being made to further simplify all aspects of this procedure, and to make it generally

available to all elementary school principals.

We are, of course, aware that numerous errors may have been incurred in the collection of the data as described in this paper. Thus, although all of our assumptions indicate that the results derived from the use of these utility functions have ratio properties, we only advise their use on an ordinal basis as a "guide" to the decision makers. Even then, we expect that the results may be counter-intuitive for some principals. However, in these latter cases, we feel (and this has been strengthened by responses in the field test) that the process of using this information will improve the principal's insight and eventual decisions.

Bibliography

1. Amor, J. P., & Dyer, J. S., "A Decision Model for Evaluating Potential Change in Instructional Programs," CSE Report No. 62, Center for the Study of Evaluation, University of California, Los Angeles, August 1970.

2. Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPortland, J., Mood, A. M., Weinfeld, F. D., and York, R. L., Equality of Educational Opportunity, U. S. Department of Health, Education and Welfare, Washington, 1966.

3. Davidson, A., & Marschak, J., "Experimental Tests of a Stochastic Decision Theory," in C. W. Churchman & P. Ratoosh (Eds.), Measurement, Definitions and Theories, John Wiley and Sons, New York, 1959.

4. _____, Suppes, S. & Siegel, S., Decision Making: An Experimental Approach, Oxford University Press, London, 1957.

5. Debreu, G., "Topological Methods in Cardinal Utility Theory," in K. J. Arrow, S. Kachin, & P. Suppes (Eds.), Mathematical Methods in the Social Sciences, 1959, Stanford University Press, Stanford, 1960.

6. De Groot, M. H., "Some Comments on the Experimental Measurement of Utility," Behavioral Science, 8 (1963), pp. 146-148.

7. Dyer, J. S., "A Procedure for the Selection of Educational Goal Areas for Emphasis," Center for the Study of Evaluation, University of California, Los Angeles, November 1971.

8. Fishburn, P. C., "Independence in Utility Theory with Whole Product Sets," Operations Research, 13, (1965), pp. 28-45.

9. _____, "Methods of Estimating Additive Utilities." Management Science, 13 (1967), pp. 435-453.

10. _____, Utility Theory for Decision Making, John Wiley and Sons, New York, 1970.

11. Hoepfner, R., Bradley, P. A., Klein, S. P., and Alkin, M. C. CSE Elementary School Evaluation KIT: Needs Assessment. Center for the Study of Evaluation, University of California, Los Angeles 1971.

12. _____, Nelken, I., Bradley, P. A., Strickland, G. P., Williams, R. C., Woolley, D. C., and Barnes, D. "Report on the Field Testing of the CSE Elementary School Evaluation KIT: Needs Assessment," CSE Report No. 70. Center for the Study of Evaluation, University of California, Los Angeles, 1971.

13. _____, Strickland, G., Stangel, G., Jansen, P., and Patalino, M. CSE Elementary School Test Evaluations, Center for the Study of Evaluation, University of California, Los Angeles, 1970.

14. Keeney, R. L., "Utility Independence and Preferences for Multiattributed Consequences," Operations Research, 19 (1971), pp. 875-893.

15. Miller, J. R., "Assessing Alternative Transportation Systems," RM-5865-DOT, The RAND Corporation, April 1969.

16. Pardee, S. F., Kirkwood, T. S., Kramer, K. L., MacCrimmon, K. R., Miller, J. R. III, Phillips, C. T., Ranstell, J. W., Smith, J. V., & Whitcomb, D. J., "Measurement and Evaluation of Transportation System Effectiveness," RM-5869-DOT, The RAND Corporation, September 1969.

17. Raiffa, H. "Preferences for Multi-Attributed Alternatives," RM-5868-DOT/RC, The RAND Corporation, April 1969.

18. Siegel, A. S., "A Method for Obtaining an Ordered Metric Scale," Psychometrika, 21 (1956), pp. 207-216.

19. Suppes, P. & Winet, M., "An Axiomization of Utility Based on the Notion of Utility Differences," Management Science, 1 (1955), pp. 259-270.