Assessing and Studying Utility Functions in Psychometric Decision Theory

Bastiaan J. Vrijhof, Gideon J. Mellenbergh, and Wulfert P. van den Brink
University of Amsterdam

In educational and industrial psychology, utility theory has been used for determining optimal decision-theoretic procedures such as optimal test cutting scores for Pass/Fail and Accept/Reject decisions. Three methods are described for empirically assessing utility functions: (1) a method for scaling utility mixtures, consisting of a true achievement or criterion level combined with the probability of passing the test or being accepted, which is applicable for determining optimal decision procedures; (2) a method for scaling the utility as a function of the true achievement or criterion level; and (3) a graphical procedure for choosing a utility function. These methods are useful for investigating the utility structure. The three methods are investigated using 30 students in a hypothetical educational Pass/Fail situation and appear to yield reliable information. Moreover, an overview of the students' utility structures is reported.

Cronbach and Gleser (1957) have introduced decision theory in psychometrics. Recently this interest has been renewed. Examples are optimal cutting scores in mastery testing (Hambleton & Novick, 1973; Huyhn, 1976), selection (Alf & Dorfman, 1967; Chuang, Chen, & Novick, 1981; Petersen, 1976), and culture-fair selection (Gross & Su, 1975; Mellenbergh & van der Linden, 1981); decision-theoretic test coefficients (van der Linden & Mellenbergh, 1978); the optimality concept (Mellenbergh & van der Linden, 1979); item selection (Mellenbergh & van der Linden, 1982) and aptitude-treatment interaction (van der Linden, 1981).

In an overview of utility measurement, Hull, Moore, and Thomas (1973) distinguish between methods that make minimal assumptions and those that make assumptions, either about the form or about the properties of the utility function. In psychometric decision theory it is mainly methods that make assumptions that have been used; examples are threshold (Hambleton & Novick, 1973), linear (van der Linden & Mellenbergh, 1977), normal ogive (Novick & Lindley, 1978), power (Huyhn, 1980), and truncated normal and extended beta (Chen & Novick, 1982) utility functions. An advantage of these functions is that they nicely match with psychometric densities as in the beta binomial or bivariate normal model. A disadvantage is, however, that it is not known if they adequately reflect an individual's utility structure. In this paper methods are described for empirically assessing an individual's own utility structure, making minimal assumptions on the utility functions. The interest is in the utility structure itself and not in the application to decision making.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 341–357
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/030341-17$2.10

Within the class of methods making minimal assumptions, Hull et al. (1973) describe, among others, rating and gambling methods; these methods are also discussed by Keeney and Raiffa (1976). Two variants of gambling methods have been used (Novick, 1980; Novick & Lindley, 1979).

In the fixed probability method the subject is offered a lottery (O_p,O_n) consisting of a preferred option O_p with probability Π and a nonpreferred option O_n with probability $(1 - \Pi)$. The subject is asked to which option O_i between O_n and O_p (s)he is indifferent between the lottery and O_i "for sure." For example, the subject is offered a lottery of winning \$100 ($O_p = 100$) with probability .75 ($\Pi = .75$) or of losing \$300 ($O_n = -300$) with probability $1 - .75 = .25$. The subject indicates that (s)he is indifferent between the lottery (100, −300, .75) and accepting \$10 for sure ($O_i = 10$). In the fixed state method the options O_p, O_n, and O_i are fixed and the subject must indicate to which probability (s)he is indifferent between the lottery (O_p, O_n, Π) and O_i for sure. For example, the subject is asked to indicate to which probability Π (s)he is indifferent between the lottery (100, −300, Π) and \$10 for sure.

In both methods the subject is indifferent between the lottery (O_p, O_n, Π) and O_i for sure; therefore, the utilities are set equal to each other:

$$U(O_i) = U(O_p,O_n,\Pi). \qquad [1]$$

Under the axioms of utility (see, e.g., Coombs, Dawes, & Tversky, 1970, chap. 5) the utility of the lottery equals its excepted value:

$$U((O_p,O_n,\Pi)) = \Pi U(O_p) + (1-\Pi)U(O_n). \qquad [2]$$

Using Equations 1 and 2, the utility of O_i is computed when the utilities of O_p and O_n are given. For example, if the utility of O_p is 1 and the utility of O_n is 0, the utility of O_i is Π; in the example $U(10) = U(100, -300, .75) = .75\ U(100) + (1-.75)\ U(-300) = .75 \times 1 + .25 \times 0 = .75$.

In this paper two rating methods and a graphical procedure making minimal assumptions on the utility functions are studied. One of the rating methods yields utility functions that can be applied in decision making, whereas the other two can only be used for describing an individual's utility structure.

Utility Functions

An essential point in decision-theoretic procedures is the specification of a loss or, equivalently, a utility function. Suppose an observed, discrete variable X ($X=0, 1, \ldots, n$) is used for making decisions on a continuous true state-of-nature variable Z. The general structure of a decision-theoretic procedure is the maximization of the expected utility, where the expectation is taken with respect to both X and Z (see, e.g., Ferguson, 1967, chap. 1):

$$E(U) = \sum_{x=0}^{n} \int_{-\infty}^{\infty} U(d(X),Z)k(X,Z)dZ, \qquad [3]$$

where $k(X,Z)$ is the joint density of X and Z; $d(X)$, the decision as a function of X; and $U(d(X),Z)$, the utility as a function of Z and the decision.

In educational and psychological measurement the observed variable X is usually a test score. The true state-of-nature variable Z is the latent trait that the test is measuring or the criterion that the test is predicting. The most important decisions based on the test score are dichotomous: Pass and Fail or Accept and Reject. Consequently, the utility function can be split into two parts and the expected utility is

$$E(U) = \sum_{x=0}^{c-1} \int_{-\infty}^{\infty} U_F(Z)k(X,Z)dZ + \sum_{x=c}^{n} \int_{-\infty}^{\infty} U_P(Z)k(X,Z)dZ, \qquad [4]$$

where $U_F(Z)$ and $U_P(Z)$ are, respectively, the utility functions for the failed (rejected) and passed (accepted) subjects, and c the cutting score on the test.

Examples of functions that make assumptions on the form are the linear (van der Linden & Mellenbergh, 1977) and normal ogive (Novick & Lindley, 1978) utility functions. Van der Linden and Mellenbergh (1977) assumed that $U_F(Z)$ is a linear decreasing function and $U_P(Z)$ a linear increasing function of Z. Novick and Lindley (1978) assumed that $U_F(Z)$ is a decumulative normal ogive and $U_P(Z)$ a cumulative normal ogive function of Z. Another example is obtained defining a cutting point d on the latent achievement (criterion) level variable Z: Examinees below d are nonmasters (nonsuited), whereas examinees equal to or above d are masters (suited). It is assumed that for examinees below d the utility does not vary with Z, and the same assumption is made for examinees equal to or above d. Equation 4 reduces to:

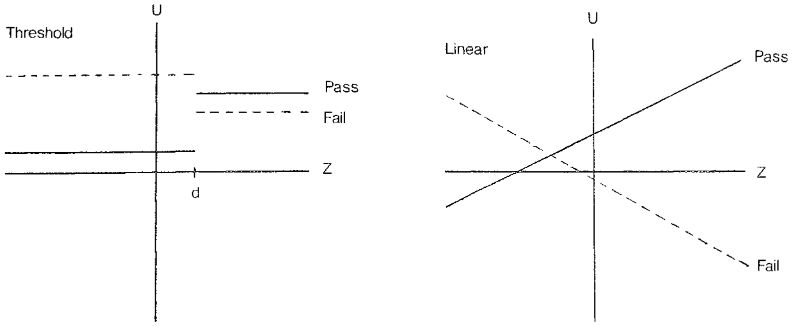
$$E(U) = U_{FL}\sum_{x=0}^{c-1}\int_{-\infty}^{d} k(X,Z)dZ + U_{FH}\sum_{x=0}^{c-1}\int_{d}^{\infty} k(X,Z)dZ + U_{PL}\sum_{x=c}^{n}\int_{-\infty}^{d} k(X,Z)dZ + U_{PH}\sum_{x=c}^{n}\int_{d}^{\infty} k(X,Z)dZ, \qquad [5]$$

where U_{FL} is the constant utility for all failed (rejected) low achievement (criterion) level examinees, U_{FH} the constant utility for all failed (rejected) high achievement (criterion) level examinees, and so on. The function is the threshold utility described by Hambleton and Novick (1973). The threshold and linear utility functions are shown in Figure 1.

Scaling Utility Function Mixtures

The pyschometric utility functions, such as threshold, linear, and normal ogive are partly a priori specified. The type of function is prescribed as threshold or linear or normal ogive. Within each type, the functions differ according to their parameter values. For example, the linear utility functions depend on their slope and intercept parameters. These parameters could be determined empirically; but without specifying the type of the function, the complete utility functions $U_P(Z)$ and $U_F(Z)$ can be scaled empirically. A very difficult problem is, however, how to scale both functions on the *same* measurement scale. It must be done by comparing the utility of, for example, a passed student with the utility of a failed student,

Figure 1
Examples of Threshold and Linear Utility Functions



given the true achievement level. For example, the student must indicate, given a true level of 60% achievement, the relation between the utility for passed and the utility for failed. A possible solution to this problem could be the combination of the two utility functions such that only one function should be scaled.

Equation 4 is written as

$$E(U) = \sum_{x=c}^{n} \int_{-\infty}^{\infty} \{U_P(Z) - U_F(Z)\} k(X,Z)\,dZ + \sum_{x=0}^{n} \int_{-\infty}^{\infty} U_F(Z) k(X,Z)\,dZ. \qquad [6]$$

Noting that the last term is a constant, it follows that maximizing $E(U)$ yields the same solution as maximizing

$$\sum_{x=c}^{n} \int_{-\infty}^{\infty} \{U_P(Z) - U_F(Z)\} k(X,Z)\,dZ = \sum_{x=c}^{n} \int_{-\infty}^{\infty} U_D(Z) k(X,Z)\,dZ, \qquad [7]$$

where $U_D(Z)$ is the difference between the two utility functions. Therefore, it is sufficient to scale the difference of the utility functions (see, for example, Chuang, Chen, & Novick, 1981).

Another combination for the two utility functions is obtained by rewriting Equation 4 as

$$\begin{aligned} E(U) &= \sum_{x=0}^{c-1} \int_{-\infty}^{\infty} U_F(Z) P(X|Z) h(Z)\,dZ + \sum_{x=c}^{n} \int_{-\infty}^{\infty} U_P(Z) P(X|Z) h(Z)\,dZ \\ &= \int_{-\infty}^{\infty} U_F(Z) P_F(Z) h(Z)\,dZ + \int_{-\infty}^{\infty} U_P(Z) P_P(Z) h(Z)\,dZ \\ &= \int_{-\infty}^{\infty} \left[U_P(Z) P_P(Z) + U_F(Z)\{1 - P_P(Z)\}\right] h(Z)\,dZ \\ &= \int_{-\infty}^{\infty} U_M(Z) h(Z)\,dZ \end{aligned} \qquad [8]$$

where

$P(X|Z)$ is the conditional distribution of X given Z,
$h(Z)$ is the marginal distribution of Z,
$P_F(Z) = P(X < c|Z)$ is the probability to fail (reject) given Z, and
$P_P(Z) = P(X \geq c|Z)$ is the probability to pass (accept) given Z.

$U_M(Z)$ is a weighted mixture of the utility functions; the weights are the conditional probabilities, respectively, to pass (accept) and to fail (reject).

The weighted mixture of the utility functions is

$$U_M(Z) = P_P(Z) U_P(Z) + \{1 - P_P(Z)\} U_F(Z). \qquad [9]$$

Comparing Equations 2 and 9 shows an analogy: $U_M(Z)$ can be considered as the expected utility of a lottery consisting of the options Pass (Accept) with probability $P_P(Z)$ and Fail (Reject) with probability $\{1 - P_p(Z)\}$. It does not make sense to offer the examinee these two options and then to ask which option between Pass (Accept) and Fail (Reject) for sure is equivalent to the lottery: There does not exist an option between Pass (Accept) and Fail (Reject). It makes sense to scale the mixture directly, however: The examinee must indicate the utility of having a certain true achievement level with a specified probability to pass (accept) and thus also a specified probability to fail (reject). In the selection situation the probability to be accepted, given the criterion level, can be estimated from empirical data. In the educational Pass/Fail situation the probability to pass the test, given a true achievement level, can be computed using the binomial distribution.

In this paper a method is described to scale the utility mixture $U_M(Z)$. The method is empirically applied and the reliability is investigated. The utility mixture does not show the form of the utility functions; therefore, two methods for assessing these functions are also studied.

Method

Assessing the Utility of the Mixture

The method for assessing the utility of the mixture is illustrated for a Pass/Fail situation existing in the psychology program of the University of Amsterdam. The examinees were students. They were instructed to imagine a situation which was hypothetical but not too far from reality: The student is preparing for an examination on psychological theories. A large domain of short answer open-ended questions is available, which are dichotomously scored correct or incorrect. The test consists of 12 items from this domain and the student passes if 6 or more items are correctly answered. A passed student receives 4 credit points, which is the equivalent of 4 weeks (160 hours) study time; a failed student does not receive credit points. The student is told (s)he needs the 4 weeks to master 50% of the total domain.

The true score continuum is split into 11 levels. Using the binomial distribution for each true score level, the probability to pass the 12-item test is computed. The following true score levels with, between parentheses, the corresponding probability to pass the 12-item test were used: 10% (.0005), 20% (.02), 30% (.12), 40% (.33), 45% (.47), 50% (.61), 55% (.74), 60% (.84), 70% (.96), 80% (.99), and 90% (.995). The percentages are presented as deviations from the 50% mastery point: −40%, −30%, ..., +40%. It is assumed that the sign of the deviation is an aid to the student's memory: A negative sign denotes that the student has studied less than 40 hours and a positive sign more than 40 hours. Moreover, the students are instructed that the study time is an increasing function of the true score level. A stimulus was a combination of a deviation from the 50% mastery point and the corresponding probability to pass the test. All possible pairs of stimuli were formed. The student indicated the preference and strength on a rating scale (Bechtel, 1976). An example is shown in Figure 2. The student was given extensive instruction containing some examples.

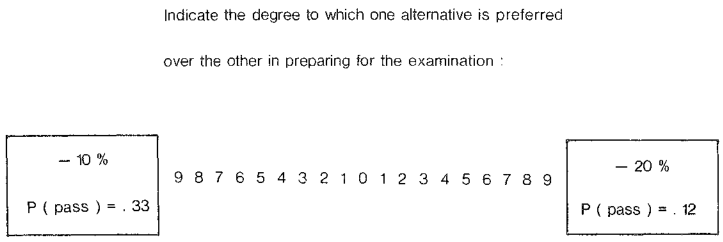
The pairs were presented in such an order that the distance between pairs containing an identical stimulus was optimal and that each stimulus was about equally often the first and the second member of the pair (Ross, 1934). Each pair was presented on a separate page of a booklet.

The rating of student i ($i = 1, 2, \ldots, N$) for the stimuli j and k is denoted d_{ijk}, which is an integer in the range from −9 to +9. The rating of the pair must be decomposed into the utilities of the two separate stimuli. Therefore, the following scaling model (Bechtel, 1976, chap. 2) was used:

$$d_{ijk} = u_{ij} - u_{ik} + y_{jk} + e_{ijk}, \qquad [10]$$

with constraints

Figure 2
Example of the Utility Mixture Method


Indicate the degree to which one alternative is preferred

over the other in preparing for the examination :

$$\sum_{j=1}^{n} u_{ij} = 0, \; i = 1, 2, \ldots, N, \qquad [11]$$

$$\sum_{k=1}^{n} y_{jk} = 0, \; j = 1, 2, \ldots., n, \qquad [12]$$

where

$$y_{kj} = -y_{jk}, \; y_{jj} = 0, \; j, k = 1, 2, \ldots., n. \qquad [13]$$

where

the parameters u_{ij} and u_{ik} are student i's utilities for stimuli and j and k,
y_{jk} is the interaction between the stimuli, and
e_{ijk} is the residual term.

The interaction term is called the unscalability, which may indicate that the separate stimuli cannot be represented on an additive utility scale. The parameters were estimated using the least squares method.

Assuming that the residuals are independently multivariate normally distributed with homogeneous variance and using the F-statistic, several null hypotheses can be tested. First, the null hypothesis that simultaneously for all students the utilities are equal is

$$H_{01} : u_{i1} = u_{i2} = \ldots. = u_{in}, \; i = 1, 2, \ldots., N. \qquad [14]$$

Second, the same null hypothesis for each student separately is

$$H_{02}^{(i)} : u_{i1} = u_{i2} = \ldots. = u_{in} \text{ for fixed } i. \qquad [15]$$

Third, the null hypothesis that the unscalability term is zero is

$$H_{03} : y_{jk} = 0, \; j, k = 1, 2, \ldots., n. \qquad [16]$$

The results of these tests are summarized in an ANOVA-like table.

Next to the statistical tests, descriptive measures of model fit can be used. From the model of Equation 10 and the least squares parameter estimates, the observed values are reproduced:

$$d'_{ijk} = \hat{u}_{ij} - \hat{u}_{ik} + \hat{y}_{jk}. \qquad [17]$$

A convenient and appropriate measure of fit for linear models, such as regression and analysis of variance models, is the product-moment correlation coefficient. The squared correlation is the percentage dependent variable variance explained by the model. Bechtel (1976, chap. 2) uses the correlation as a measure of fit for the linear scaling model Equation 10. The product-moment correlation R_u between the observed d_{ijk} and the reproduced d'_{ijk}, computed over all data, is an overall measure of fit for model Equation 10. Under the null hypothesis that the unscalability parameter equals zero, the model of Equation 10 reduces to

$$d_{ijk} = u_{ij} - u_{ik} + e_{ijk}. \qquad [18]$$

Using this model and the least squares parameter estimates, the observed values are reproduced by

$$d''_{ijk} = \hat{u}_{ij} - \hat{u}_{ik}. \qquad [19]$$

The product-moment correlation R_s between the observed d_{ijk} and the reproduced d''_{ijk}, computed over all data, is a measure of fit for the model of Equation 18. If the difference $(R_u - R_s)$ is small, the unscalability parameter y is negligible; if at the same time R_s is high, the scalability model of Equation 18 is an adequate overall description of the data. Thus, it makes sense to represent the stimuli on an additive utility scale.

Although the model of Equation 18 may be an adequate overall description, it is possible that the model does not fit for particular individuals. Therefore, the correlations R_{si} ($i = 1, 2, \ldots, N$) were computed between d_{ijk} and d''_{ijk} for each separate student. Inspection of the results shows which students fail to fit the scalability model Equation 18.

The measures R_s and R_u are easily computed using the ANOVA-like table (Bechtel, 1976, p. 27). It is noted that for the descriptive measures no distributional assumptions are made. The least squares estimates and the correlations are computed without distributional assumptions. The measures are therefore also valid when the distributional assumptions necessary for the F tests are violated.

Assessing the Utilities

For assessing the utility functions $U_F(Z)$ and $U_P(Z)$, two methods were used: Comrey's constant sum method (Torgerson, 1958, p. 105) and a graphical procedure.

In the constant sum method the same 11 true score levels were used, but these were presented without the probabilities of passing the test. All possible pairs of the true achievement level percentages were formed. First, the student was told that (s)he had passed the test and that (s)he must indicate for each stimulus of the pair whether the utility is positive or negative. Then, the student was asked to split 100 points into two parts according to the utility ratio of the two stimuli. An example is shown in Figure 3. The student has indicated that the utility for both −30% and −10% is negative, and (s)he has split the 100 points into 50 and 50 points, which means that the utility of the two stimuli is equal. Second, the same student was told that (s)he has failed the test, and the whole procedure was repeated.

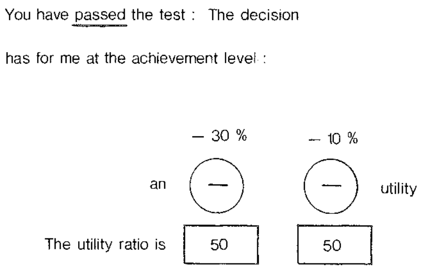
The method yields the ratio of the absolute values of the utilities and the sign for each separate utility. The model for the ration w_{ijk} of the absolute values of the utilities of stimuli j and k is considered to be multiplicative:

$$w_{ijk} = \frac{|u_{ij}|}{|u_{ik}|} y_{jk} e_{ijk} = \frac{\exp(u^*_{ij})}{\exp(u^*_{ik})} \exp(y^*_{jk}) \exp(e^*_{ijk})$$

$$= \exp(u^*_{ij} - u^*_{ik} + y^*_{jk} + e^*_{ijk}) . \qquad [20]$$

Taking the natural logarithm, the model of Equation 10 follows for the log ratios, and the same methods as described above were applied. The procedure yields least squares estimates for u^*_{ij} ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n$). Taking the antilog of $\hat{u}^*_{ij}$ yields the estimate of the absolute value of the utility. The

Figure 3
Example of Constant Sum Method



You have passed the test : The decision

has for me at the achievement level :

The utility ratio is 50 50

sign of the utility is derived from the student's pluses and minuses assigned to the separate stimuli. The assumption that the residuals e_{ijk} are independently multivariate normally distributed is unrealistic in the model of Equation 20. Therefore, it is better to rely on the descriptive measures rather than on the *F* statistics.

For the graphical procedure a book with about 1,800 utility function graphs is prepared. The design was of the funnel type: On the first page the student chose from a series of different threshold functions. Depending on the choice, the student was presented a second page with figures that were specifications of the choice of the previous page, and so on. In the procedure the utility functions change gradually from threshold to linear and from linear to convex and concave functions. In each step the student could stick to his (her) last choice, which means that (s)he stopped and was not forced to go on. If the student stopped, (s)he was presented related figures from previous steps to check or correct his (her) choice. The procedure was applied twice, both for the Pass situation and the Fail situation.

Procedures

The methods were used with 30 second-year students from the Department of Psychology. Each student was individually tested by the first author. For each student five tasks were prepared in the same sequence: The mixture preference rating, the constant sum judgment for Pass and for Fail, and the graphical judgment for Pass and for Fail. The sequence of the pairs in the paired comparison tasks was alternated: For half of the students one sequence was used; and for the other half, the reversed sequence. The time needed for task administration was about 2 hours and students were paid for their participation.

Results

Utility Mixtures

The results are summarized in Table 1. The *F* statistics show that for all students simultaneously the utilities differ from each other. For each separate student the utilities differ significantly; but for some students, such as Nos. 7 and 29, the *F* statistics are relatively low. The *F* statistic shows that the unscalability parameters differ from zero. The difference between the correlations—R_u = .85 and R_s = .81—is not very large, and the addition of the unscalability parameter does not yield a much better data description. The model of Equation 18 appears to be an acceptable descriptive model for practical purposes. The correlations per student show that for some students, such as Nos. 9 and 13, the model is not very adequate.

Inspecting the utility function graphs shows two types. The first is an increasing function; a typical example is given in Figure 4(a). This type is found for Students Nos. 2, 8, 9, 10, 22, and 27; their utility mixture generally increases with true achievement level. The second type is found for the other students, except No. 29, and is exemplified in Figure 4(b); their utility mixture increases to a maximum and then decreases. Intuitively both types make sense: Students from the first type want to study as much as possible to obtain the highest possible score, whereas the students from the second type want to study until their own optimal point.

Utility Functions

The statistics and descriptive measures are computed for the log ratios, i.e., on the logarithmic scale. The functions are graphed on the utility scale itself.

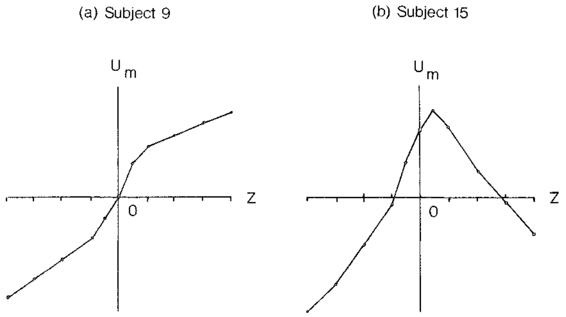
Table 1

F-Statistics and Correlations for Preference Strength Ratings of the Utility of the Mixtures

Source of Variation		SS	DF	MSS	F		R_{si}
Utilities							
All Subjects		55689	300	185.66	20.47	R_u=.85	
Per Subject:	1	2336	10	233.61	25.75		.89
	2	2300	10	229.98	25.35		.73
	3	1454	10	145.39	16.03		.80
	4	485	10	48.51	5.35		.82
	5	2588	10	258.77	28.53		.60
	6	1955	10	195.48	21.55		.80
	7	260	10	25.98	2.86		.79
	8	1874	10	187.37	20.66		.80
	9	3014	10	301.44	33.23		.49
	10	3240	10	323.99	35.72		-
	11	1748	10	174.76	19.27		.80
	12	569	10	56.94	6.28		.86
	13	2859	10	285.88	31.52		.55
	14	1546	10	154.63	17.05		.94
	15	2216	10	221.61	24.43		.87
	16	2110	10	211.00	23.26		.74
	17	1585	10	158.45	17.47		.86
	18	1848	10	184.84	20.38		.72
	19	2171	10	217.11	23.93		.87
	20	2051	10	205.10	22.61		.87
	21	1772	10	177.24	19.54		.90
	22	1840	10	183.99	20.28		.69
	23	1225	10	122.54	13.51		.83
	24	1657	10	165.69	18.27		.92
	25	2234	10	223.39	24.63		.65
	26	2022	10	202.21	22.29		.68
	27	1639	10	163.91	18.07		.76
	28	3095	10	309.51	34.12		.86
	29	259	10	25.87	2.85		.60
	30	1746	10	174.60	19.25		.78
Unscalability		3458	45	76.85	8.47	R_s=.81	
Error		11838	1305	9.07			
Total		70994	1650				

Note. All F-statistics are significant at the 1% level. For Subject 10 the correlation was undefined.

Figure 4
Examples of Empirical Utility Functions, Mixture Preference Rating Method



The *F* statistics and measures of fit for the Pass situation are reported in Table 2. The unscalability parameter is not substantial: The difference between the correlations—R_u = .88 and R_s = .87—is very small and the *F* statistic is not significant. The model of Equation 18 is, except for Student No. 3, a reasonably good description of the data. The graphs of the utility functions of most students are generally increasing functions; examples are given in Figure 5 (a,b). Eight students show a deviant utility function; examples are given in Figure 5 (c, d).

The statistics for the Fail situation are reported in Table 3. The results are the same as for the Pass situation: The unscalability is negligible and, except for Student No. 3, the model of Equation 18 is a good data description. The utility graphs for all but one student are generally decreasing functions; examples are given in Figure 6.

Figure 5
Examples of Empirical Utility Functions,
Constant Sum Method, for the Pass Situation

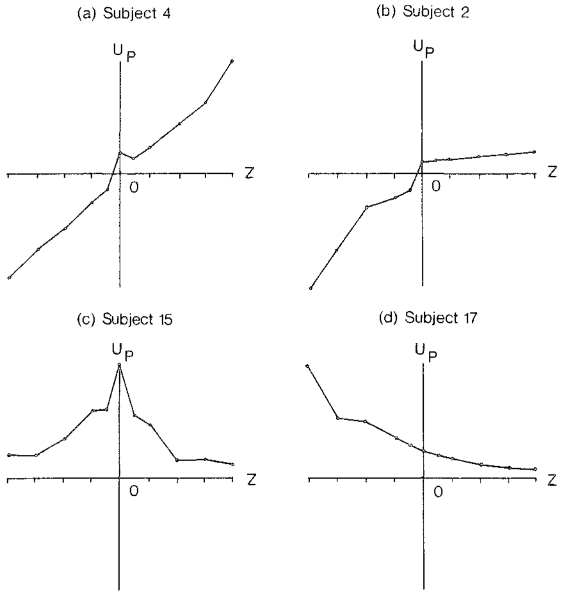


Table 2
F-Statistics and Correlations for Logarithmic Transformed Utility Ratios in the Pass Situation

Source of Variation		SS	DF	MSS	F		R_{si}
Ln-Utilities							
All Subjects		2205.3	300	7.35	14.44	R_u=.88	
Per Subject	1	69.1	10	6.91	13.58		.81
	2	55.9	10	5.59	10.98		.87
	3	13.3	10	1.33	2.62		.51
	4	56.3	10	5.63	11.06		.96
	5	354.7	10	35.47	69.69		.87
	6	154.9	10	15.49	30.44		.78
	7	12.0	10	1.20	2.35		.68
	8	21.1	10	2.11	4.14		.85
	9	24.7	10	2.47	4.86		.81
	10	59.3	10	5.93	11.65		.91
	11	109.3	10	10.93	21.48		.70
	12	61.6	10	6.16	12.11		.88
	13	205.5	10	20.55	40.38		.83
	14	28.9	10	2.89	5.69		.91
	15	56.7	10	5.67	11.13		.83
	16	57.6	10	5.76	11.31		.91
	17	70.5	10	7.05	13.85		.92
	18	34.1	10	3.41	6.69		.83
	19	119.4	10	11.94	23.47		.82
	20	78.3	10	7.83	15.39		.73
	21	14.1	10	1.41	2.78		.74
	22	84.2	10	8.42	16.54		.90
	23	45.6	10	4.56	8.96		.70
	24	136.0	10	13.60	26.72		.91
	25	102.3	10	10.23	20.11		.77
	26	44.5	10	4.45	8.74		.85
	27	27.0	10	2.70	5.30		.91
	28	37.4	10	3.74	7.34		.93
	29	34.7	10	3.47	6.81		.77
	30	36.3	10	3.63	7.14		.83
Unscalability		22.8	45	.51	.99	R_s=.87	
Error		664.2	1305	.51			
Total		2892.2	1650				

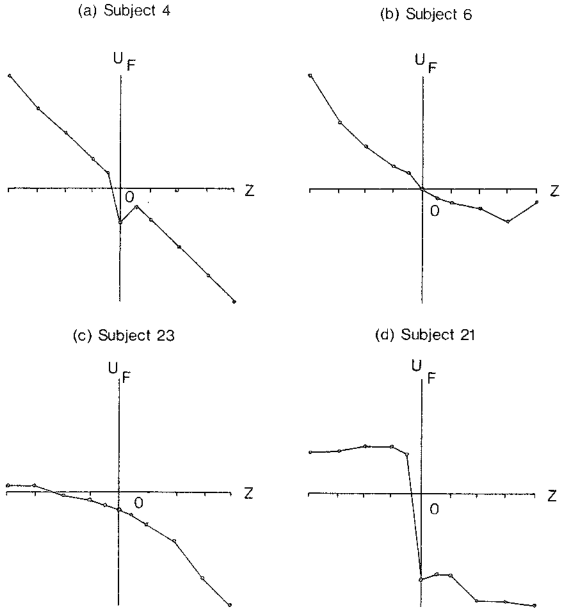
Note. All F-Statistics, except for the unscalability, are significant at the 1% level.

Table 3
F-Statistics and Correlations for Logarithmic
Transformed Utility Ratios in the Fail Situation

Source of Variation		SS	DF	MSS	F		R_{si}
Ln-Utilities							
All Subjects		2828.9	300	9.43	26.79	R_u=.91	
Per Subject	1	112.2	10	11.22	31.89		.82
	2	120.0	10	12.00	34.09		.90
	3	19.5	10	1.95	5.54		.54
	4	53.5	10	5.35	15.21		.95
	5	476.9	10	47.69	135.53		.91
	6	88.0	10	8.80	25.01		.84
	7	17.1	10	1.71	4.86		.79
	8	25.6	10	2.56	7.26		.79
	9	56.5	10	5.65	16.05		.95
	10	51.4	10	5.14	14.60		.97
	11	202.3	10	20.23	57.49		.92
	12	141.5	10	14.15	40.20		.90
	13	97.5	10	9.75	27.71		.87
	14	25.6	10	2.56	7.28		.93
	15	119.5	10	11.95	33.96		.60
	16	46.6	10	4.66	13.24		.87
	17	72.4	10	7.24	20.58		.96
	18	51.2	10	5.12	14.55		.85
	19	130.1	10	13.01	36.96		.90
	20	273.9	10	27.39	77.84		.93
	21	22.2	10	2.22	6.31		.89
	22	96.8	10	9.68	27.50		.92
	23	135.7	10	13.57	38.56		.88
	24	132.0	10	13.20	37.51		.90
	25	113.6	10	11.36	32.28		.82
	26	43.6	10	4.36	12.38		.83
	27	16.2	10	1.62	4.60		.94
	28	45.2	10	4.52	12.85		.88
	29	31.6	10	3.16	8.99		.80
	30	10.6	10	1.06	3.02		.61
Unscalability		18.6	45	.41	1.17	R_s=.89	
Error		459.3	1305	.35			
Total		3306.8	1650				

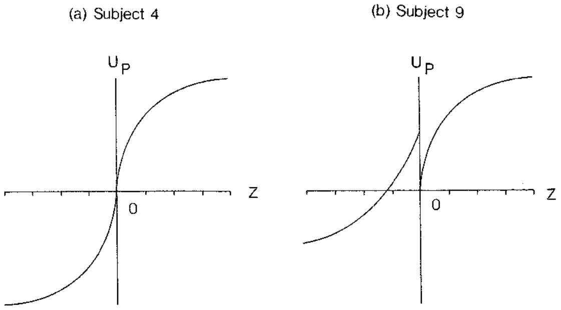
Note. All F-Statistics, except for the unscalability , are significant at the 1% level.

Figure 6
Examples of Empirical Utility Functions, Constant Sum Method, for the Fail Situation



Finally, the results of the graphical procedure are considered. In the Pass situation six students chose a continuous increasing curve; for an example, see Figure 7(a). Nine students chose discontinuous, increasing curves; see Figure 7(b). One student shows a threshold function indicated by two horizontal lines. Eleven students show a combination of a horizontal or increasing straight line or curve followed by a horizontal or increasing straight line or curve. In the Fail situation, seven students show a continuous, decreasing curve and five a discontinuous, decreasing curve. The remaining 18 students show a combination; none shows a threshold function.

Figure 7
Examples of Empirical Utility Functions, Graphical Method, for the Pass Situation



As a check on the consistency, the results of the graphical and constant sum methods were compared. For the Pass situation the utility functions from the two methods were compared for the true achievement level below 50%. Three judgments were made. First, are the two functions in the same positions, below or above the horizontal axis? For example, comparing Figures 5(a) and 7(a) shows that this is the case for Student No. 4. Second, are the two functions in the same direction: increasing, decreasing, or flat? Figures 5(a) and 7(a) show that this is also true for Student No. 4. Third, are the two functions of the same curvature: convex, concave, or straight? The figures show that this is not the case for Student No. 4.

The three judgments were also made for true achievement levels above 50%. The whole procedure was repeated for the Fail situation. The results are reported in Table 4. The table shows high agreement in position and direction but rather low agreement in curvature; the utility functions of the constant sum method are generally straight lines, whereas those of the graphical procedure are more often curved.

Discussion and Conclusions

Both the constant sum and the graphical procedure are reliable methods for assessing the conditional utility functions. The form of the functions shows that threshold functions are not adequate description of the students' utility structure. For some students linear utility functions are reasonable approximations of their utility structure, but many students deviate from linearity.

The constant sum and graphical procedures are suited for studying the form of the utility functions. They are less suited for determining optimal decisions. A serious problem is that the utility function for the Pass situation is not scaled on the same measurement scale as the one for the Fail situation. As remarked before, the experimental task of scaling the functions with respect to each other is rather impractical for a subject. Even if the students could perform this task, a conceptual problem remains for them: They must scale the conditional utilities as a function of the true achievement level *without* knowledge of the probability to pass the test. The abstract task of scaling utilities, without knowledge of the probability to pass, questions the validity of the task. The scaling of the utility mixture, where the student knows the probability to pass, seems better suited for this purpose.

The data show that the mixtures are reliably scaled. The additive utility model of Equation 18 is approximately valid, although the unscalability parameter differs significantly from zero. For practical purposes the method could be used for assessing an individual's or group's utility mixture. For example, in determining optimal cutting scores, Equation 8 is maximized as a function of the cutting score c. In the experiment the students assessed the utility mixture $U_M(Z)$ for a given cutting score $c = 6$ on a 12-item test. Consequently, using Equation 8, $E(U)$ can only be estimated for the cutting score $c = 6$. For determining the optimal cutting score, however, $E(U)$ must be estimated for all possible cutting scores: $c = 0, 1, 2, \ldots, n$, and for each cutting score the probability to pass (accept)—$P_P(Z)$ in Equation 8—has another value. To solve this problem the utility mixtures must be scaled for at least one other cutting score, say c'. From Equation 9 follows that the utility of the mixture equals:

$$U'_M(Z) = P'_P(Z)U_P(Z) + \{1-P'_P(Z)\}U_F(Z), \qquad [21]$$

where $P'_P(Z)$ is the probability to pass (accept) for the cutting score c' as a function of the true achievement (criterion) level and $U'_M(Z)$ is the corresponding utility of the mixture. Note that for each new cutting score the probability to pass (accept) usually has another value. For a given value Z Equations 9 and 21 are two equations with two unknown parameters, i.e., $U_P(Z)$ and $U_F(Z)$; $P'_P(Z)$ is computed using the binomial distribution (educational situation) or empirical data (selection situation). Solving these equations for all values of Z used in the experiment yields the functions $U_P(Z)$ and $U_F(Z)$. Using these functions in Equation 9 yields $U_M(Z)$ for all possible cutting scores $c = 0, 1, 2, \ldots, n$, which can be used in Equation

Table 4
Agreement(+) and Disagreement(-) Utility Functions
from the Constant Sum and Graphical Procedure

	Pass						Fail					
	Position		Direction		Curvature		Position		Direction		Curvature	
	Bel.	Ab.	Bel.	Ab.	Bel.	Ab.	Bel.	Ab.	Bel.	Ab.	Bel.	Ab.
	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
Subject												
1	+	+	+	+	+	–	+	+	+	+	+	+
2	+	+	+	+	–	–	+	–	+	+	–	+
3	+	+	+	+	–	+	–	+	+	+	–	+
4	+	+	+	+	–	–	+	+	+	+	–	–
5	+	+	+	+	+	–	+	+	+	+	–	–
6	+	+	+	+	–	+	+	+	+	+	–	+
7	–	+	+	+	+	–	–	–	+	–	–	–
8	+	+	–	+	–	–	+	+	+	+	+	–
9	–	+	+	+	+	–	+	+	+	+	–	–
10	+	+	+	+	–	–	+	+	+	+	–	–
11	–	+	+	+	–	–	–	+	–	–	–	–
12	+	+	+	+	–	+	+	+	+	+	–	+
13	+	+	–	–	–	–	+	+	+	–	–	–
14	+	+	+	–	+	–	+	+	+	+	+	+
15	+	+	+	+	+	+	+	+	+	+	–	–
16	+	+	+	+	–	–	+	+	+	+	–	–
17	+	+	+	+	–	+	+	+	+	+	+	+
18	+	+	+	+	+	–	+	+	+	+	+	–
19	+	+	+	+	–	–	+	+	+	+	–	–
20	–	+	+	+	–	–	+	+	+	+	–	+
21	+	+	+	+	–	–	+	+	+	+	–	+
22	+	+	+	+	–	+	+	+	+	+	–	–
23	+	+	+	+	+	+	+	+	+	+	+	–
24	+	+	+	+	+	–	+	+	+	+	+	–
25	–	+	–	–	–	–	–	+	+	+	+	+
26	+	+	+	+	+	+	+	+	+	+	+	+
27	–	+	+	+	+	+	+	+	+	–	–	–
28	+	+	+	–	–	–	+	+	+	+	–	+
29	–	+	–	+	–	–	+	+	–	+	–	–
30	+	–	+	–	+	–	–	–	–	–	–	–
Total	23	29	26	25	12	9	25	27	27	25	9	12

8 to estimate $E(U)$ for all possible cutting scores. The optimal cutting score is the value of c for which the expected utility is maximal. In the example the application to the decision problem cannot be demonstrated because the utility mixture has only been scaled for the value $c = 6$. In applications to decision problems the mixture must be scaled for at least two different cutting scores.

An interesting point is whether the constant sum and graphical procedure yield the same results as the mixture. For the data it is impossible to make the comparison because the mixture is only scaled for one cutting score $c = 6$. Moreover, the mixture cannot be reproduced from the separate utility functions because the mixture is measured on an interval level, whereas the separate functions are ratios. If, however, the mixture is scaled for at least two cutting scores, the separate utility functions $U_P(Z)$ and $U_F(Z)$ can be derived. These can be compared to the corresponding functions from the constant sum and graphical procedure.

In the experiment a finite number of values of the true achievement (criterion) level Z were used that yielded a finite number of values for the mixture $U_M(Z)$. To these values a polynomial can be fitted:

$$U_M(Z) = b_0 + b_1 Z + b_2 Z^2 + \ldots + b_g Z^g . \qquad [22]$$

Substituting Equation 22 into Equation 8 yields

$$E(U) = b_0 + b_1 E(Z) + b_2 E(Z^2) + \ldots + b_g E(Z^g) , \qquad [23]$$

where $E(Z)$, $E(Z^2)$, …, $E(Z^g)$ are the moments of the distribution $h(Z)$. Therefore, for distribution functions with higher order moments equal to zero, Equation 23 is simplified. For example, if $h(Z)$ is a normal distribution of criterion scores, Equation 23 reduces to

$$E(U) = b_0 + b_1 E(Z) + b_2 E(Z^2) = b_0 + b_1 \mu + b_2 (\sigma^2 + \mu^2) , \qquad [24]$$

where μ and σ^2 are the mean and variance of h(Z), which are estimated from a sample. The coefficients b_0, b_1, and b_2 are determined from the experiment scaling the mixture $U_M(Z)$. It is emphasized that for each $c = 0, 1, 2, \ldots, n$ the coefficients have different values. As remarked before, however, for computing the coefficients for all cutting scores $c = 0, 1, 2, \ldots, n$ it is sufficient to scale $U_M(Z)$ for only two cutting scores.

In the selection situation $h(Z)$ is the distribution of the observed criterion scores which can be estimated from empirical data. In the educational Pass/Fail situation $h(Z)$ is the distribution of the latent scores. In the beta-binomial model it is assumed that $h(Z)$ has a beta distribution with parameters that can be estimated from the observed distribution (Lord & Novick, 1968, chap. 23). Moreover, in the model it is assumed that each examinee answers a *different* set of n randomly selected items from a large item domain. In the usual educational situation all examinees answer the *same n* items. It is not known how robust the beta-binomial model is against this assumption violation (see van den Brink, 1982).

References

Alf, E. F., & Dorfman, D. D. The classification of individuals into two criterion groups on the basis of a discontinuous payoff function. *Psychometrika*, 1967, *32*, 115–123.

Bechtel, G. G. *Multidimensional preference scaling*. 's-Gravenhage, The Netherlands: Mouton, 1976.

Chen, J. J., & Novick, M. R. On the use of a cumulative distribution as a utility function in educational or employment selection. *Journal of Educational Statistics*, 1982, *7*, 19–35.

Chuang, D. T., Chen, J. J., & Novick, M. R. Theory and practice for the use of cut scores for personnel decisions. *Journal of Educational Statistics*, 1981, *6*, 129–152.

Coombs, C. H., Dawes, R. M., & Tversky, A. *Mathematical psychology: An elementary introduction*. Englewood Cliffs NJ: Prentice-Hall Inc., 1970.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1957.

Ferguson, T. S. *Mathematical statistics: A decision-theoretic approach*. New York: Academic Press, 1967.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology*, 1975, *60*, 345–351.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159–170.

Hull, J. C., Moore, P. G., & Thomas, H. Utility and its measurement. *Journal of the Royal Statistical Society, Series A*, 1973, *136*, 226–247.

Huyhn, H. Statistical considerations of mastery scores. *Psychometrika*, 1976, *41*, 65–79.

Huyhn, H. A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika*, 1980, *45*, 167–182.

Keeney, R. L., & Raiffa, H. *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley, 1976.

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

Mellenbergh, G. J., & van der Linden, W. J. The internal and external optimality of decisions based on tests. *Applied Psychological Measurement*, 1979, *3*, 257–273.

Mellenbergh, G. J., & van der Linden, W. J. The linear utility model for optimal selection. *Psychometrika*, 1981, *46*, 283–293.

Mellenbergh, G. J., & van der Linden, W. J. Selecting items for criterion-referenced tests. *Evaluation in education: An international review series*, 1982, *6*, 171–183.

Novick, M. R. Statistics as psychometrics. *Psychometrika*, 1980, *45*, 411–424.

Novick, M. R., & Lindley, D. V. The use of more realistic utility functions in educational applications. *Journal of Educational Measurement*, 1978, *15*, 181–191.

Novick, M. R., & Lindley, D. V. Fixed-state assessment of utility functions. *Journal of the American Statistical Association*, 1979, *74*, 306–311.

Petersen, N. S. An expected utility model for "optimal" selection. *Journal of Educational Statistics*, 1976, *1*, 333–358.

Ross, R. T. Optimum orders for the presentation of pairs in the method of paired comparisons. *Journal of Educational Psychology*, 1934, *25*, 375–382.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Van den Brink, W. P. Binomial test models for domain-referenced testing. *Evaluation in education: An international review series*, 1982, *6*, 160–170.

Van der Linden, W. J. Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. *Psychometrika*, 1981, *46*, 257–274.

Van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, 1977, *1*, 593–599.

Van der Linden, W. J., & Mellenbergh, G. J. Coefficients for tests from a decision-theoretic point of view. *Applied Psychological Measurement*, 1978, *2*, 119–134.

Acknowledgments

The authors thank Ed Broeze, Henk Kelderman, Wim J. van der Linden, Pieter Vijn, and Michel Zwarts for their comments, and Erna Bot and Meredith Iqbal for typing the manuscript.

Author's Address

Send requests for reprints or further information to Gideon J. Mellenbergh, Psychologisch Laboratorium, Universiteit van Amsterdam, Weesperplein 8, 1018 XA Amsterdam, The Netherlands.