COWO

Passing scores on domain referenced tests: an improved decision-theoretic methodology for optimization.

Ben Wilbrink

may 1980

[aangeboden aan Applied Psychological Measurement. Retourbrieven zijn zoekgeraakt zowel bij de verhuizing van het COWO naar de Turfmarkt, als bij de verhuizing van mijn correspondent namens APM, Wim van der Linden, naar Twente. Daardoor was ik ondertussen zo ver verwijderd van het onderwerp dat ik geen definitieve versie heb gemaakt.
Deze ingezonden versie lijdt onder beroerd Engels, ik had de tekst niet eerst op het Engels laten corrigeren.
De tekst is te geserreerd opgezet, waardoor de pointe van een aantal belangrijke vernieuwingen de lezer ongetwijfeld zullen ontsnappen. Het helpt om eerst de conclusies te lezen., en daarna de voorbeelden diagonaal door te nemen.
Dit rapport past in de traditie van werken met wiskundige functies als kandidaten voor nutsfuncties, met telkens een specifieke wiskundige oplossing voor het 'optimum.' Dat schiet niet op, en blijft het nadeel van kunstmatighoud houden. Achteraf gezien is simpel: in ieder nieuw specifiek geval gaat het om dezelfde bewerking, dat is de weging van een voorspellende verdeling over de doelvariabele met de nutsfunctie over die doelvariabele. Die weging mas mathematisch allerlei interessante vormen hebben, maar rekentechnisch komt het altijd op hetzelfde neer en is er dus geen behoefte aan al die specifieke analytische oplissingen. Dit inzicht ontbreekt in de vigerende literatuur, evenals in dit rapport. Het is een inzicht dat kennelijk alleen met veel moeite is te verwerven, moete die ik heb gedaan in jarenlang werken aan de programmering van een algemeen toetsmodel.]

Centre for Research on Higher Education (C.O.W.O.)
University of Amsterdam
Oude Turfmarkt 149 / 1012 GC Amsterdam, Netherlands.

PRE-PUBLICATION; please do not quote without permission from the author.

[In de in 2002 toegevoegde voetnoten geef ik commentaar op de theoretische lijn zoals die in dit rapport in 1980 is ontwikkeld, en die deels veel te dicht bij die van het 'vigerende model' is gebleven. Op legio punten van ondergeschikt belang, die ik vandaag de dag zeker niet meer zo zou uitwerken, laat ik commentaar achterwege.]

abstract

A decision analytic methodology is presented for setting the passing score on domain referenced tests. Distinguishing feature of this methodology is the explicit treatment of expected remediation effects. These effects not longer being considered implicit in the utility assignments results in more transparent utility structures. Attention is given to several classes of utility functions. The mathematical development, based on extensive form analysis, is rather simple compared to normal form analysis that minimizes the Bayes risk. It is not assumed that a 'mastery score' is set on the domain score dimension. In this methodology the terminal[1] utility functions for the 'pass' and 'remediate' alternatives will in general not intersect. For a non-trivial passing score to exist it is sufficient that expected terminal utility functions intersect. The resemblance of the presented methodology to Cronbach and Snow's (1977) Aptitude Treatment Interaction methodology is summarily pointed out.

In domain referenced testing interest lies in the degree of competence attained by a particular student, or the group of students that followed the particular instructional unit. However, on the end-of-unit test decisions will have to be made either to retain students for extra study time or remediation, or pass them to the next instructional unit. This kind of decision does not imply that a mastery score or cutting score is set on the underlying true score dimension; only on the test score dimension a *passing score* must be established, below which students are retained for some remediation.[2]

The most promising approach to the problem of locating the passing score that is optimal in some specified sense is the decision theoretic one (Glass, 1978). The way in which this approach usually is implemented, however, suffers from undue restriction to the mathematically complicated 'normal form analysis', as in Huynh (1976), Peters (1976), Van der Linden (1980). Mathematically more tractable is the 'extensive form analysis', being preferred for that reason by Raiffa and Schlaifer (1961) and, in domain referenced testing, by Davis, Hickman and Novick (1973). Both forms of analysis will result in the same optimal decisions, as Raiffa and Schlaifer (1961) showed, accordingly I will use the simpler one.[3]

The decision's optimality depends on variables called 'states of nature', about whom the decision maker will have some, by no means perfect, information. In domain referenced testing the state of nature will be the domain score,[4] [5] the proportion of items in the domain the student would answer correctly when given the opportunity. Usually there is only one state of nature involved, as e.g. in medical diagnosis, and the unknown state of nature will determine the outcome of the decision made. Among other things a probability model is needed connecting available information to the state of nature; using it as a weighting function on the utilities the decision maker has assigned to the state of nature, expected terminal utilities will be determined, and the diagnosis having the higher expected value is chosen. That is a fine approach to use on the medical problem, but is it also the right approach to the special problem of locating the optimal passing score in domain referenced testing?

A unique feature of this special decision problem is that one of the alternative treatments is meant to influence the very state of nature the decision is based on. The goal of remediation is to further the proficiency of the student in answering items from the domain, and the expected results of remediation will influence the level at which the passing score is set. It will clarify the situation to distinguish the state of nature 'domain score (reached)' *before* and *after* remediation. I will refer to specific scores as π and t respectively, indicating the difference, although the dimension is of course the same. Two probability models will now be needed: the first probability model connects the end-of-unit observed score X to the domain score P; the second connects the same observed score to the domain score after remediation T. A second test, after remediation, is only used to estimate the domain scores after remediation. Should after remediation again a decision be made to retain or pass students, the same decision analysis is to be repeated; this complication may be avoided by deciding to always pass students after remediation.[6] [7]

The utility function on domain scores[8] will have to be specified by the decision maker; several classes of utility functions to choose from will be considered. Remark that the same state of nature is relevant to both decision alternatives, and that accordingly *the utility function on domain scores will be the same under both decision alternatives.* Remediation involves certain costs in teacher and student time, use of extra facilities, extra testing, but these costs do not influence the utility assigned to domain scores: they will be brought into the analysis separately.[9]

The decision situation being properly analyzed,[10] the decision theoretic apparatus can now be brought to bear on it. Remark that nowhere a 'mastery score' or the like will be needed, meeting the one and only objection of Glass (1978) to the decision theoretic approach. To avoid every suggestion that somewhere, somehow, a 'mastery score' should be specified, I will speak of domain referenced testing, instead of criterion referenced testing. The to-be-specified passing score need not have any surplus meaning over and above what is implied in the utility and cost specifications.

Linear utility

The linear utility function is the simplest function to handle in decision analysis, and it will be the perfect vehicle to demonstrate the decision theoretic methodology to be used in order to locate the passing score optimally.[11] Threshold utility may seem even simpler, but it involves the choice of the threshold score, so it assumes part of the problem already being solved.[12]

Motivation for the use of linear utility could be found in the special character of the learning material: when a lot of rather disconnected facts are to be learned, for example, this utility function will be one of the best to use. In a medical context Raiffa and Schlaifer (1962) use the example of choice between a new drug and the hitherto used one, where the relevant state of nature is proportion π of patients cured. The linear utility function may be chosen as $u = \pi$;

The specification of the scale of utility is free up to a positive linear transformation, i.e. the optimal decision will not be affected by this kind of transformation. So the linear utility function on domain score (reached)[13] may be chosen as

$$u = \pi. \qquad [1]$$

It serves no purpose to use instead a more general formulation like $a\pi + b$, so equation 1 is the best choice to make.

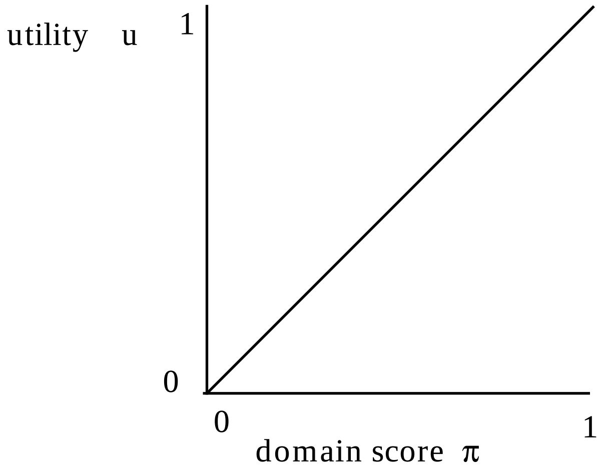

Figure 1. Linear utility function u = π on domain scores.

Figure 1 pictures this nice function. If questions from the domain may be answered correctly by guessing the answer or the correct alternative, the same function may still be used: it is not necessary to evaluate the 'chance level' as having zero utility. Extending the linear utility function below chance level need not bother us, for expected domain scores will never be this low.

Having assigned this linear utility we can construct the *terminal* utility functions corresponding to the decision alternatives 'pass' and 'remediate'. It is important to carefully distinguish these two kinds of functions; later we will meet a third kind of function: *expected* terminal utility. The 'pass' decision involves no other utilities or costs than those specified in equation 1, so *a passed student results in the utility corresponding to his domain score.*[14] The terminal utility function for the pass decision[15] is accordingly equal to the utility function in equation 1:

$$u_t\ (p, \pi) = \pi, \qquad [2]$$

where the suffix t denotes a terminal utility, and p denotes the 'pass' decision.

It might be objected that a passed student with a low domain score will run a grave risk to fail his next unit because of inadequate preparation. This objection boils down to choosing rather an external criterion, like success on next unit(s), than the internal criterion of domain score reached; in this paper I will use this internal criterion.[16]

For retained students going through remediation the interest is still in the domain score, now of course the domain score reached after remediation. It is however the same dimension, the domain not having changed in the meantime, so without the costs of remediation the terminal utility function for the decision 'remediatel would also be equal to the function specified in equation 1. But there is a cost involved: the cost of setting up remedial instruction, the cost of extra time spent by teacher and student, the cost of retesting, the possible loss of motivation when the student with high domain score is retained. In general there will be a negative relation between domain score and extent of remediation that is necessary; there will possibly be a positive relation between domain score and loss of motivation because of (unnecessary) remediation; costs of retesting are not related to domain score. These considerations suggest that *constant costs* c might be a good approximation.[17] And remarking that extreme scores will not affect the passing score, it is sufficient to assume the costs constant in the neighborhood of the passing score.

So a retained student results in the utility corresponding to his domain score, less the constant cost c. [18] The terminal utility function for the decision remediate is:

$$u_t\ (r, t) = t - c, \qquad [3]$$

where r denotes the decision remediate.

Instead of the constancy assumption it is of course possible to specify the costs as a function of domain score, where the choice can be made from the several classes of utility functions presented in this paper, among them the linear ones. In the extensive form analysis only the expected cost given observed score $X = x$ will be needed, so it is possible even with variable costs to work with a constant value for given $X = x$ (see also Raiffa and Schlaifer 1961 p. 17).

In the medical drug example the new drug had some nasty side effects over and above the side effects the old drug had; these side effects being evaluated on the already specified utility scale as c, Raiffa and Schlaifer give as terminal utility functions for new drug n and old drug o: $u_t\ (n, \pi) = \pi$ and $u_t\ (o, \pi) = \pi - c$.[19]

The cost c is to be evaluated on the now established utility scale, e.g. as a proportion of the cost of the 'regular' instructional activities of this unit.[20] Evaluating c as .25, the situation is as depicted in figure 2.

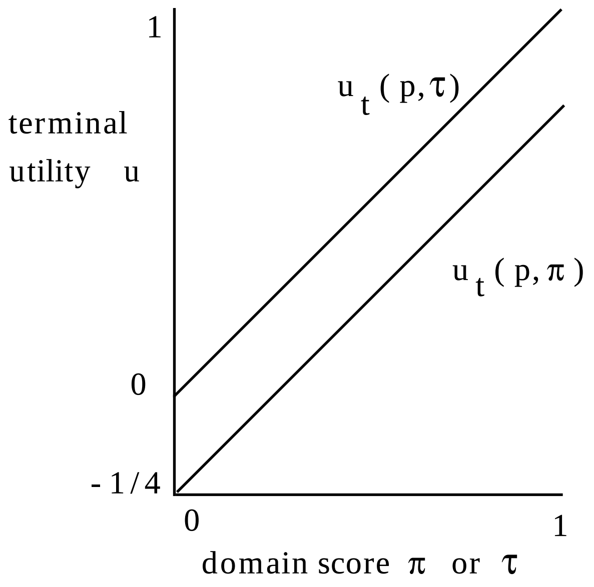

Figure 2. Terminal utility functions u_t (p, π) = π for the 'pass' decision, and u_t (r, t) = t + c for the 'remediate' decision; c is expected cost of remediation given X = x.

Figure 2 is remarkable, there being no intersection of the terminal utility functions, not even a disordinal one. It must be realized that figure 2 involves only the terminal utilities: the effect that remediation has on the domain score is not yet taken account of. This effect enters the analysis through the conditional p.d.f.'s on domain score before and after remediation, see figure 3. [21]

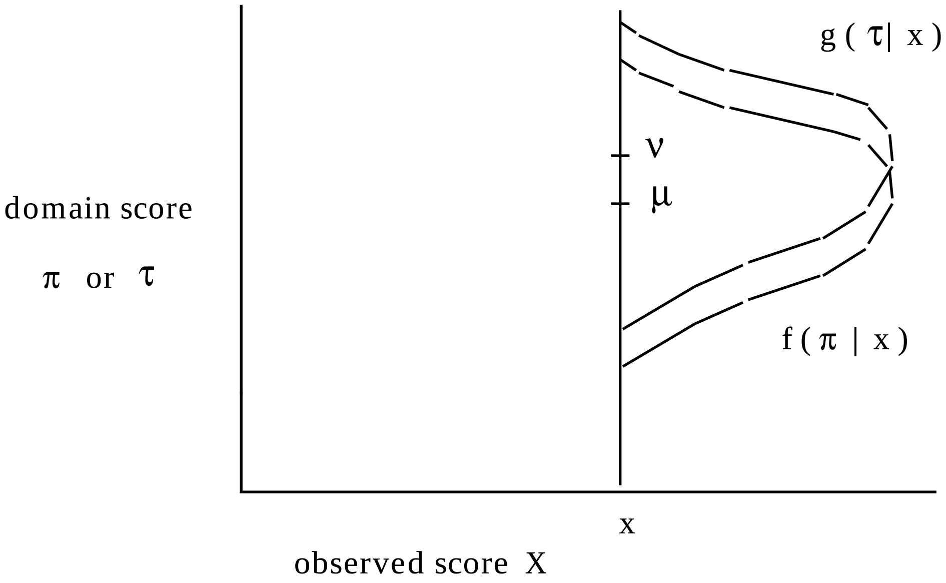

Figure 3. The conditional probability density functions on domain score, before remediation f (π | x), and after remediation g (t | x).

Let f (π |x) be the p.d.f. of domain scores (before remediation) conditional on the observed score X = x. Then the *expected terminal utility* for this decision,
$E_{\pi|x}\ u_t\ (p, \pi)$, is obtained by summing the products of utility and probability for every P = π:

$$E_{\pi | x}\ u_t\ (p, \pi) = \int_0^1 \pi\ f\ (\pi \mid x)\ d\pi = \mu, \qquad [4]$$

where the suffix $\pi|x$ denotes that the expectation is taken with respect to f ($\pi|$x), and p is the conditional expected domain score (a suffix on μ to denote that it is conditional on X = x is suppressed, to increase readability). [22] Assuming the regression of domain score on observed score to be linear, [23] μ can be estimated using:

$$\mu = r_{xx'}\ .\ m_x\ / n + (l - r_{xx'})\ .\ m_x\ / n\ , \qquad [5]$$

where r is an appropriate 'reliability' coefficient, and m_x the mean observed score.

Considering the decision 'remediate', let $g(t|x)$ be the p.d.f. of domain scores *after* remediation given X = x on the *same* end-of-unit test. Remark that in this model a *prediction* of the domain score after remediation is involved. The expected terminal utility for this decision is:

$$E_{\tau|x}\ u_t(p, t) = \int_0^1 (c + \tau)\ g(\tau \mid x)\, d\tau = \nu + c, \qquad [6]$$

where n is the expected domain score after remediation, given X = x.

A *validation study* on the effect of remediation is necessary, using an unselected group of students that is given remediation, to estimate n. For this estimate E (y | x) may be used, Y being the observed score on the end-of-remediation test, because $E(t|x) = E(y - \varepsilon|x) = E(y|x)$ assuming $E(\varepsilon|x) = 0$, where ε is the error score random variable. Assuming regression of T or Y on to be linear we have:

$$\nu = E(y \mid x) = m_y / n + (x - m_x) . r_{xy}\, s_y / s_x\, n , \qquad [7]$$

where r_{xy} is the observed correlation between X and Y, and s_x and s_y are the standard deviates.

The decision rule is: select the alternative having the greater expected terminal utility. Comparing equations 4 and 6 the student is passed if:

$$c \geq \nu - \mu, \qquad [8]$$[24]

or if the expected cost of remediation is greater than the expected gain in domain score. By this rule the optimum passing score is set: by calculating the preferred alternative for every X = x, or analytically. The optimal passing score q has the unique characteristic that:

$$E_{\pi|q}\ u_t(p, \pi) = E_{\tau|q}\ u_t(p, \tau) \qquad [9]$$

disregarding the discreteness of X. It is the observed score where the decision maker is indifferent as to either retain or pass students having this observed score.

Calculation of expected utilities for some trial values of X will give the turnover point, and the optimal passing score as the observed score equal to or just above this point. This assumes there will be only one 'optimal' passing score, an assumption that may be made in this kind of educational assessment, an assumption also that will be inherent in some probabilistic models like the beta-binomial one that will later be given.

The general 'solution' for the optimal passing score is, alternatively, given as the observed score that makes equation 8 into an equality. This is the same result as obtained by minimizing the Bayes risk in 'normal form analysis', as I will show now.

Normal form analysis

I will show here the normal form analysis, as used by Van der Linden, in the case of linear utility. The results of this paragraph will not be used in the sequel. In normal form analysis the *Bayes risk* is minimalized, the Bayes risk being, crudely expressed, the total expected loss over all examinees for a given passing score. 'Loss' is equal to the negative of the terminal utilities as considered in the preceding paragraph. With the linear utility structure from the preceding paragraph the Bayes risk in setting the passing score at t is:

$$B(t) = \sum_{x=t}^{n} \int_0^1 -\pi \ f(x \mid \pi)\, f(\pi)\, d\pi \ + \sum_{x=0}^{t-1} \int_0^1 (-\tau + c) \ g(x \mid \tau)\, f(\tau)\, d\tau, \quad [10]$$

which may be rewritten as:

$$B(t) = \sum_{x=t}^{n} \int_0^1 -\pi \ f(\pi \mid x)\, h(x)\, d\pi \ + \sum_{x=0}^{t-1} \int_0^1 (-\tau + c) \ g(\tau \mid x)\, h(x)\, d\tau. \quad [11]$$

This equation simplifies to:

$$B(t) = \sum_{x=t}^{n} -\mu \ h(x) \ + \sum_{x=0}^{t-1} (-\nu + c) \ h(x). \quad [12]$$

Adding terms to the first sum and subtracting them from the second this becomes:

$$B(t) = \sum_{x=0}^{n} -\mu \ h(x) \ + \sum_{x=0}^{t-1} (\mu - \nu + c) \ h(x). \quad [13]$$

The first term in equation 13 does not depend on t, so B (t) is minimized when the second term in equation 13 is minimized. Assuming that there will be only one optimal cutting score, this may be found by stepwise increasing the 'trial' passing score t: when the sum stops decreasing, the corresponding t of the last step will be the optimal passing score q. It is easy to see that this will happen when (μ - ν + c) gets positive for the first time, or when c is at least equal to μ - ν, the result also obtained in the preceding paragraph by the simpler extensive form analysis. Comparing both forms it is seen that normal form analysis uses equations 12 and 13 on top of the equations that are used by extensive form analysis, while the unconditional expected loss integrals in equations 10 and 11 are more complicated than the conditional expected terminal utility integrals in extensive form analysis.

The beta-binomial model

Using linear utility it is not necessary to specify the p.d.f.'s f (π | x) and g (τ | x), the optimal decision depending only on the expected values. Using other utility functions it will either be necessary, or make things a lot easier, to specify these probability models. Natural candidates when dealing with achievement tests are the beta functions arising in the beta-binomial model for observed scores X. There the beta-binomial p.d.f. βb(a, b, n) is fitted on the observed score distribution, e.g. by estimating the parameters a and b by the method of moments, using equations 14 and 15:

$$b = \frac{s^2 - m\,(n - m)}{m - n\,s^2 \,/\, (\,n - m\,)} \qquad [14]$$

$$a = b\,m \,/\, (\,n - m\,) \qquad [15]$$

where n is the number of testitems, m the mean observed score, and s the observed standard deviation.

The function itself will not be needed, although a graphic plot of it may be made to inspect the fit. The function is calculated term-by-term, or approximated (see Raiffa and Schlaifer 1961). Its function is:

$$h\,(\,x\,) = \beta b(\,a, b, n\,) = \binom{n}{x} B^{-1}(\,a, b\,)\, B\,(\,a + x, b + n - x\,), \qquad [16]^{25}$$

where B (a, b) = (a - l)! (b - l)! / (a + b - l)!.

Assuming linear regression of domain score on observed score, the probability density function of domain scores will be the beta function:

$$f (\pi) = \beta (a, b) = B^{-1} (a, b) \pi^{a - l} (1 - \pi)^{b - l}. \qquad [17]$$

The observed-score distribution given $\Pi = \pi$ will under this model be the binomial density function:

$$h (x \mid \pi) = \binom{n}{x} \pi^{x} (1 - \pi)^{n - x}. \qquad [18]$$

However, it is $f (\pi \mid x)$ we are interested in. Using a well-known relation it can be shown that:

$$f (\pi \mid x) = f (\pi) \; h (x \mid \pi) \; / \; h (x) =$$

$$B^{-1} (a + x, b + n - x) \pi^{a + x - 1} (1 - \pi)^{b + n - x - 1} \qquad [19]$$

Equation 19 is equal to the beta function $\beta (a + x, b + n - x)$.

Mean and variance of $\beta b(a, b, n)$ are:

$$\frac{n a}{a + b} \quad \text{and} \quad \frac{n a b (a + b + n)}{(a + b)^{2} (a + b + 1)} \qquad [20]$$

Mean and variance of the beta density $\beta (a, b)$ are

$$\frac{a}{a + b} \quad \text{and} \quad \frac{a b}{(a + b)^{2} (a + b + 1)} . \qquad [21]$$

The r^{th} moment about zero of a (a ,b) is:

$$\mu_r' = B^{-1} (a, b) B(a + r, b). \qquad [22]$$

To my knowledge there is no joint beta-binomial distribution available, so the functional form of $g (\tau \mid x)$ will have to be assumed as for example being normal. This assumption will simplify calculations. In the beta-binomial model this conditional density will only be beta when the underlying beta p.d.f.'s are identical. Further assuming homoscedascity, the variance of this conditional normal p.d.f. is:

$$\sigma_{\tau}{}^{2} \ (1 - \rho_{x\tau}{}^{2}), \qquad [22]$$

where the correlation between X and T is obtained using:

$$\rho_{x\tau}{}^{2} = r_{xy}{}^{2} \ / \ \rho_{yy'} , \qquad [24]$$

see e.g. Lord and Novick (1968) formula 3.9.7. The reliability, or correlation between randomly parallel forms, or alternatively the Kuder Richardson formula 21 reliability (see Lord and Novick 1968 formula 23.6.14 that, after using equation 20 and correcting the deviant notation in this book, is equal to equation 25) is in the beta-binomial model:

$$\rho_{yy'} = n \ / \ (n + c + d), \qquad [25]$$

where c and d are the parameters of the βb (c, d, n) p.d.f. fitted on the observed scores Y.

A useful property in evaluating expected terminal utilities is, also for higher powers of π:

$$\int \pi \, \beta (a, b) \ d\pi \ = \ (a \ / \ (a + b)) \ \int \beta (a + 1, b) \ d\pi \ . \qquad [26J$$

Definite integrals of the beta p.d.f. can be evaluated from K. Pearson's *Tables of the incomplete beta function*, Biometrika, London 1934, or from tables of the cumulative binomial function, e.g. *Tables of the cumulative binomial probability distribution* by The staff of the Computation Laboratory, Cambridge, Mass., Harvard University Press, 1955. The cumulative binomial may be used because of the following relation with the cumulative beta:

$$F_\beta (\gamma; a, b) = \begin{cases} 1 - F_b (a ; \gamma, a + b - 1) & \gamma \leq 0.5 \\ F_b (b; l - \gamma, a + b - 1) & \gamma \geq 0.5 \end{cases} \qquad [27]$$

where F (β) is the cumulative beta distribution function, and F (b) the cumulative binomial (number of successes, success parameter y or 1 - y, and a + b - l the number of trials).

One remark concerning the beta-binomial fit must be made. If the number of testees is small, the estimators may not be very stable, and some care must be exercized in interpreting the results. However, a passing score will have to be set, also when available (statistical) information is scant. So the model may be used even with small numbers of testees on which estimates are to be based: using the model will give the greater expected benefit, a principle of decision analysis (Raiffa and Keeney 1976, Lindley 1971) already clearly formulated by Simon (1943, 1977). When the βb fit is bad, or equations 14 and 15 give a negative result, other approximations may be tried, like the joint normal p.d.f.

Threshold utility

Considerably more complex than linear utility is the threshold utility case: the evaluation of expected terminal utilities is more involved, and characteristically the threshold γ on the domain score dimension must be made available first. Usually this threshold is assumed to be the ‘mastery score,’ distinguishing ‘mastery’ and ‘non-mastery’ on the particular learning tasks involved. Such a dogmatic view is quite unnecessary: threshold utility may be used as approximation to an ogive-shaped utility function. There is, however, not much merit in using this approximation, because using that ogive shaped utility function, like the cubic to be treated later, may be mathematically less involved. Threshold utility being extensively used, however, discussion can't be omitted here.

The threshold, while definitely not a ‘minimally acceptable’ level of domain score, may be viewed as the preferred level of domain score, or shorter:

the *preferred score*. Higher domain scores are of course evaluated as having higher utility, but the meaning of the preferred score is that the decision maker will be perfectly satisfied should his instruction and assesment strategy result in domain scores in the neighborhood of his preferred score. A rating technique may be used to locate this preferred score γ, see e.g. Siegel (1957) or Becker and Siegel (1962). The rating technique would in fact result in the utility function on domain scores, so techniques given by Keeney and Raiffa (1976) can also profitably be used. In this way the choice of threshold will also be based on utility assignments, the location of threshold γ being determined as the point where the function has its steepest ascent, or its second derivative is equal to zero.

Threshold utility may be chosen to be zero for domain scores below γ, and one for higher scores, using the fact that positive linear transformations of the utility scale will not affect the optimal decision, so:

$$u = \begin{cases} 0 & \text{for} \quad \pi \text{ or } \tau < \gamma \\ 1 & \text{for} \quad \pi \text{ or } \tau \geq \gamma \end{cases} \qquad [28]$$

The terminal utility function for the ‘pass’ decision is equal to equation 28, no other utilities or costs being involved:

$$u_t\,(\,p, \pi\,) = \begin{cases} 0 & \text{for} \quad \pi < \gamma \\ 1 & \text{for} \quad \pi \geq \gamma \end{cases} \qquad [29]$$

The cost of remediation being c, the terminal utility function for the decision ‘remediate’ is:

$$u_t\,(\,p, \pi\,) = \begin{cases} 0 & \text{for} \quad \tau < \gamma \\ 1 & \text{for} \quad \tau \geq \gamma\,. \end{cases} \qquad [30]$$

The expected terminal utilities are:

$$E_{\pi\,|\,x}\ u_t\,(\,p,\pi\,)\ =\int_{\gamma}^{1} f\,(\,\pi\ |\ x\,)\ d\pi\ =\ P_x \qquad [31]$$

and

$$E_{\tau\,|\,x}\ u_t\,(\,r,\tau\,)\ =\ \int_{0}^{\gamma} -c\ \ g\,(\,\tau\ |\ x\,)\ d\tau\ +\ \int_{\gamma}^{1}\,(\,1-c\,)\ \ g\,(\,\tau\ |\ x\,)\ d\tau$$

$$= -c\,(\,1-Q_x)\ +\ (\,1-c\,)\,Q_x\ =\ Q_x - c, \qquad [32]$$

where P_x and Q_x denote definite integrals, to be evaluated by using tables mentioned in the preceding paragraph when the beta-binomial model is used, or the table of the normal distribution function when the joint normal model is used.

The optimal passing score q is the observed score X = x that makes equations 31 en 32 equal (disregarding the discrete character of x, which will lead to the smallest score just above the analytically determined ‘optimal’), or where:

$$Q_x - P_x\ =\ c. \qquad [33]$$

Of course, P_x and Q_x denote the probability of a domain score above the threshold, given X = x, before and after remediation.

Piecewise utility

The technique used in threshold utility analysis may also profitably be used with piecewise utility, i.e. utility functions made up of parts of functions. For example:

$$\pi\ /\ \gamma \qquad \text{for} \qquad \pi\ <\ \gamma$$

$$u = \begin{cases} & \\ 1 & \text{for} \quad \pi \geq \gamma . \end{cases} \qquad [34]$$

where γ again denotes the preferred score.
In the evaluation of the resulting integrals for expected terminal utility, and using the beta-binomial model, the property given as equation 26 can be used. The utility function in equation 34 is a composition of two linear functions, as also is the case with threshold utility. Of course pieces of more flexible utility functions may in the same way be used to match the preferences of the decision maker as close as possible with the mathematical functions to be used in the analysis.

Quadratic and cubic utility

Increased flexibility in utility functions is to be found with exponential functions, treated in the next paragraph, and quadratic and cubic functions. The general algebraic form of the quadratic function is:

$$A + B\,\pi + C\,\pi^2. \qquad [35]$$

Again, this will also be the terminal utility function under the alternative ‘pass,’ and after being lowered by the expected cost c (for students having observed score x) it will also be the terminal utility function under the alternative ‘remediate.’

As a terminal utility function equation 35 has the nice property that it results in expected terminal utilities consisting of the constant A, and the first and second moment about zero of the p.d.f. involved.

Remembering that the second moment equals the sum of the variance and the squared mean, it follows that:

$$E_{\pi\,|\,x}\; u_t\,(\,p,\pi\,) \;=\int_0^1 (\,A + B\,\pi + C\,\pi^2\,)\; f\,(\,\pi \mid x\,)\;\; d\pi$$

$$= A + B\,\mu + C\,(\,\sigma_{\pi}^{2} + \mu^{2}\,), \qquad [36]$$

and:

$$E_{\tau\,|\,x}\; u_{t}\,(\,r, \tau\,) \;=\; A + B\,\nu + C\,(\,\sigma_{\tau}^{2} + \nu^{2}\,) + c\,. \qquad [32]$$

Adding a cubic term D π^3 gives the decision maker the opportunity to use an ogive like function to represent his utility assignments. In the expected terminal utilities this results in the added term μ_3', or in the βb model:

$$\mu_3' = \frac{B (a + 3, b)}{B (a, b)} D = D \frac{a (a + 1) (a + 2)}{(a + b) (a + b + 1) (a + b + 2)} , \qquad [38]$$

using equation 22. Now this and other yet to be treated classes of utility functions must be fitted to some known points of the *utility curve*. These points, or the whole curve, may be obtained using the already mentioned rating technique, the ‘regular’ techniques as given by Keeney and Raiffa (1976), or special techniques like the fixed state (lottery technique) utility assessment suggested by Novick and Lindley (1979). An interactive program for utility assessment is available on the Computer Assisted Data Analysis Monitor, Isaacs and Novick (1978).

Some points of the utility curve having been determined, a cubic function may be determined using four points, checking the consistency of the calculated function with other points, and correcting until a satisfactory fit is reached. With the cubic function the point of inflection may be used, corresponding to the preferred score γ, having the property that the second derivative is zero, and f (0) = 0 and f (1) = 1 may be chosen (as a first approximation, to be changed if necessary for a better fit in the critical ranges of the domain scores).

Exponential utility

With utility functions of the exponential type, expected terminal utility can be calculated using the exponential transform $T_x (s)$:

$$T_x (s) = E (e^{-sx}) = \int e^{-sx} f (x) dx. \qquad [39]$$

Keeney and Raiffa (1976, § 4.9.6) give a list of these transforms for some common probability density functions. For the beta density β (a, b):

$$E (e^{-sx}) = \sum_{j=0}^{\infty} \frac{(a + j)! \; (a + b)! \; (-s)^{j}}{a! \, (a + b + j)! \; j !} \qquad [40]$$

this is the confluent hypergeometric function with arguments a, a + b, and -s. The function converges for $|S| < 1$. It is tabulated in Abramowitz and Stegun (1964). Johnson and Kotz (1969 p. 9) give an approximate formula:

$$E (e^{-sx}) = (a + b - l)! \, e^{-0.5\,s} \{-0.5\,s(b - a) \}^{-0.25 \; -0.5\;a \; -0.5\;b} \cos \{ 2 \{ -0.5\,s (b - a) \}^{0.5} - 0.5\,\pi (a + b -0.5) \} \qquad [41]$$

for a large and b and s fixed.

For the binomial density function with parameter π on n observations the exponential transform is:

$$E (e^{-sx}) = (\pi e^{-s} + 1 - \pi)^{n} . \qquad [42]$$

For the normal distribution $N (\mu, \sigma^{2})$:

$$E (e^{-sx}) = \exp (-s \mu + s^{2} \sigma^{2} / 2). \qquad [43]$$

Exponential functions, or sums of exponential functions, give one considerable latitude for utility specification, e.g. the (negative) utility specification on costs of (remedial) instruction.

Normal distribution utility

The use of probability density functions and probability distribution functions as utility functions is promoted by Novick and Lindley (1978), the mathematical development being given by Lindley (1976). For the passing score problem the ogive like distribution functions are particularly appropriate, the inflexion point again corresponding to the preferred score. The reason to use a distribution function resides in the easy evaluation of the expected terminal utility integrals when using for example the normal distribution function in combination with a normal p.d.f. as probability model. In the sequel I will follow the exposition of Novick and Lindley (1978).

Using a probability distribution function as utility function on domain scores Π, it may be said that the utility of $\Pi = \pi$ is equal to the probability that a random quantity, Θ, say, is smaller than π, or p ($\Theta \leq \pi$). But π is also random, having a specified p.d.f. Now the beauty of this is that the expected terminal utility, when terminal utility is equal to the utility function, is the probability that one random quantity is less than another: p ($\Theta < \Pi$), or that the difference $\Theta - \Pi$ (or the quotient $\Theta \,/\, \Pi$) is smaller than zero. Choosing normal distribution and normal density functions, the fact may be used that the difference $\theta - \pi$ has the normal distribution also.

Fitting a standardized normal ogive to one's utility curve, with mean γ (because the mean corresponds to the inflection point, and therefore to the preferred score γ) and standard deviation σ, this function is written:

$$u = \Phi ((\pi - \gamma) \,/\, \sigma), \qquad [44]$$

where Φ denotes the standardized normal ogive. When equation 44 is also the *terminal* utility function, as it may be for the ‘pass’ decision, the result for the expected terminal utility is:

$$E_{\pi | x} \, u_t (p, \pi) = E_{\pi | x} \, u = \Phi \frac{\mu - \gamma}{(\sigma_\pi^2 + \sigma^2)^{0.5}} \qquad [45]$$

where f (π | x) is N (μ, σ_π^2). This expected value may be obtained from a normal distribution table. For the decision ‘remediate’ involving cost c:

$$E_{\tau | x} \, u_t (r, \tau) = -c + \Phi \frac{\nu - \gamma}{(\sigma_\tau^2 + \sigma^2)^{0.5}} \qquad [46]$$

The normal density will in many cases be a good approximation. When the observed score distributions are skew, a normalizing transformation may be used on the data, e.g. Novick and Jackson (1974) par. 10.1. When already the beta-binomial model has been used, the obtained beta functions can be replaced by normal p.d.f.'s with the same mean and standard deviation, when neither of the parameters is smaller than 10.

Remark that, when σ is chosen great, this normal distribution utility function will approach the threshold utility function.

Novick and Lindley (1978) suggest several other possibilities in the choice of utilities using this class of functions. For example, when the conditional p.d.f.'s are beta, a member of the class of beta cumulative distribution functions may be chosen as utility function: the required computation of the difference of two beta variables is available on the CADA Monitor (Isaacs and Novick, 1978).

Numerical examples

Linear utility. Suppose the validation study, using tests with 20 items, results in end-of-unit test reliability $\rho (x, x') = 0.50$, $m = 12$, $s (x) = 3$; end-of-remediation test $y = 16$, $s (y) = 2$; and $r (x, y) = .20$.
Then, using equation 5, $\mu = 0.50 x / 20 + (1 - 0.50) 12 / 20 = 0.30 + 0.025 x$.
Using equation 7, $\nu = 16 / 20 + (x - 12) 0.2 (2 / 3) / 20 = 0.72 + .0067 x$.
Thus $\nu - \mu = 0.42 - .00183 x$.
The cost $c = 0.25$, so when $x \geq 10$ is the expected terminal utility of remediation positive.
The optimal passing score is accordingly set at 10.

Threshold utility and the beta-binomial model. Let the validation study, using 30 item tests, result in $x = 24.33$, $s (x) = 2.462$, $y = 25.92$, $s (y) = 2.111$, and $r (x, y) = 0.20$.
Using equations 14 and 15 the following beta-binomial p.d.f.'s are fitted: $h (x) = \beta b (73, 17, 30)$ and $g (y) = \beta b (95, 15, 30)$. The reliabilities (correlation between alternate forms randomly chosen from the domain) are,

using equation 25, $\rho (x, x') = 0.30$ and $\rho (y, y') = 0.27$.
Let the treshold be $\gamma = .8$, as determined by a suitable technique as mentioned in the text. Then the expected terminal utility of the 'pass' decision is according to equations 31 and 27:

$$P (x) = \int_{0.8}^{1} \beta (73 + x, 120)\, d\pi = \sum_{45 - x}^{120} bi (0.2, 120).$$

It remains to determine Q (x). Using equation 24 $\rho^2 (x, \tau) = 0.040 / .27 = 0.15$.
Then the variance of $g (\tau \mid x)$ may be taken as equation 23, using equation 21 for $\sigma^2 (\tau)$, giving $(95 * 15 / 110^2 * 111) * (1 - 0.15) = 0.000902$.
The mean of $g (\tau \mid x)$ is $E (\tau \mid x) = E (y \mid x)$, and assuming regression of Y on x linear, this is given by equation 7 as
$25.92 / 30 + (x - 24.33) * 0.2 (2.111 / 2.462) / 30 = 0.725 + 0.0057 x = \nu$.

Now choose a trial value x = 20, then $E (\tau \mid x) = 0.839$. The value of P (20) is read from a table of the cumulative binomial to be 0.4457. The value of Q (20) is read from a table of the standardized cumulative normal distribution function, entered at
$z = (\nu - \gamma) / \sigma (\tau \mid x) (0.725 + 0.114 - 0.8) / 0.03 = 1.3$, so Q (20) = 0.9032.
Now Q (20)- P (20) = 0.46, rather greater than the cost c, these being 0.25 in this example. Now choose x = 22, that results in Q (22) - P (22) = 0.24, so the expected terminal utility of remediation is here less than the cost of remediation. The optimal passing score is accordingly set at x = 22.

C*ubic utility*. Let the decision maker have fitted the cubic function $-0.87 \pi + 3.2 \pi^2 - 1.33\pi^3$ to his utility curve. For the ‘pass’ decision the expected terminal utility is according to the sum of equations 36 and 38, working with the beta-binomial model:
$- 0.87 \mu + 3.2 (\sigma^2 (\pi \mid x) + \mu^2) - 1.33 * (73 + x) (74 + x) (75 + x) /120 * 121 * 122$.
Evaluating this for the same data as used in the preceding example, again using the βb model, μ and $\sigma^2 (\pi \mid x)$ are the mean and variance of
$f (\pi \mid x) \ \beta (a + x, b + n - x)$;

using equation 21 $\mu = (73 + x) / 120$, and $\sigma^2 (\pi \mid x) = (73 + x) (45 - x) / 120 * 121$. For $x = 22$, $\mu = 0.792$ and $\sigma^2 (\pi \mid x) = 0.001254$, making the expected terminal utility for the ‘pass’ decision -0.689 + 2.011 - 0.664 = 0.658.

For the 'remediation' decision $g (\tau \mid x)$ may be taken normal with mean $\nu = 0.725 + 0.0057 x$ and variance 0.0009, as derived in the preceding example.

Now the third moment about zero ν_3' is needed:

$\nu_3' = \nu_3 + 3 \nu_1' \nu_2 + \nu_1'^3 = 3 \nu \sigma^2 + \nu^3$ because $\nu_1' = \nu$, $\nu_2 = \sigma^{2,}$ and ν_3 vanishes for symmetrical densities.

When $x = 22$, $\nu = 0.850$ and $\nu_3' = 0.616$.

Taking the cost of remediation in this example to be $c = .10$, the expected terminal utility of the ‘remediate’ decision is, using equation 36 and the here developed ν_3': $-0.87 * 0.850 + 3.2 (0.0009 + 0.850^2) - 1.33 * 0.616 - 0.10 = 0.656$.

This is the smaller expected terminal utility, suggesting that $x = 21$ is the optimal passing score.

The calculations for $x = 21$ result in expected terminal utility for the ‘pass’ decision 0.641, for the ‘remediate’ decision 0.647.

Quadratic cost. Let the decision maker have fitted the quadratic function $0.5 - 0.94 \pi + 0.55 \pi^2$ to what he estimates to be the cost of remediation for students entering remediation having realized already the domain score π. Be careful, in calculating the expected terminal cost of remediation, to weight the quadratic cost function by $f (\pi \mid x)$ and *not* by $g (\tau \mid x)$. Expected terminal cost is, using the data from the preceding example, for $x = 22$ equal to $0.5 - 0.94 * 0.792 + 0.55(0.001254 + 0.792^2) = 0.101$. For $x = 21$ the result is 0.102.

Normal distribution utility. Let the normal distribution corresponding to N (0.8, 0.01) be chosen as utility function. Let $m = 24$, $s (x) = 2.5$, and $\rho (x, x') = 0.30$. The variance $\sigma^2 (\pi) = \rho (x, x') s^2 (x) / n^2 = 0.30 \text{ x } 6.25 / 900 = 0.00208$.

The variance of $f (\pi \mid x)$ is, using equation 23, $\sigma^2 (\pi \mid x)$

$= \sigma^2 (\pi) \; (1 - \rho^2 (x, \pi)) = \sigma^2 (\pi) (1 - \rho (x, x')) = 0.00208 * 0.70 = .00146$.
The mean of $f (\pi \mid x)$ is $\mu = 0.30 x / 30 + 0.70 m / 30 = 0.01 x + 0.568$, using equation 5.
Let $f (\pi \mid x)$ then be normal $N (0.01 x + 0.568, 0.00146)$.
Evaluating first at $x = 22$, $f (\pi \mid 22) = N (0.788, 0.00146)$.
Now using equation 45:
$\Phi \{(0.788 - 0.8) / (0.00146 + 0.01)^{0.5} \} = \Phi (-0.012 / 0.1071) = \Phi (-0.112)$. Entering the table of the standardized normal distribution at $z = -0.112$ this gives the expected terminal utility of the ‘pass’ decision at $x = 22$ as 0.457.

Let $m_y = 26$, $s (y) = 2$, $\rho (y, y') = 0.25$, and $r (x, y) = 0.20$.
Then $\sigma^2(\tau) = 0.25 x 4/900 = 0.00111$, taking number of test items again $n = 30$. Now using equation 24: $\rho^2 (x, \tau) = 0.22/0.25 = 0.16$.
The variance $\sigma^2(\tau \mid x) = \sigma^2(\tau) (1 - \rho^2 (x, \tau)) = 0.00111 * 0.84 = 0.00093$.
The mean of $g(\tau \mid x)$ is, using equation $\nu = 26 / 30 + (x - 24) * 0.2 * (2 / 2.5) * 30 = 0.0053 x + 0.739$.
Let $g(\tau \mid x)$ then be normal $N(0.0053 x + 0.739, 0.00093)$.
So $g(\tau \mid 22)$ is $N (0.856, 0.00093)$.
Using equation 45: $\Phi \{(0.856 - 0.8) / (0.00093 + 0.01) 1/21$
$= \Phi (0.056 / 0.1054) = \Phi (0.531)$.
Taking cost of remediation $c = 0.25$, and entering the table of the standardized normal distribution at $z = 0.531$, this results in the expected terminal utility of the 'remediate' decision being $0.702 - 0.25 = 0.452$, slightly less than the expected terminal utility of the 'pass' decision.
For $x = 21$ expected terminal utilities are 0.419 ('pass') and 0.432 ('remediatel), so $x = 21$ is the optimal passing score.

Discussion and conclusions

A transparent methodology for the setting of passing scores on domain referenced tests has been presented, resulting in simple and elegant solutions. This tranparence was made possible by use of the ‘extensive form analysis,’ involving only the evaluation of expected terminal utility given X = x. To my knowledge this method was earlier used in domain referenced testing only by Davis, Hickman and Novick (1973). The 'standard' approach uses the 'normal form analysis' that minimizes the Bayes risk (for an overview see Van der Linden 1980). However, I demonstrated that this normal form analysis can be viewed as involving the same extensive form analysis and on top of that some surplus mathematical exercises: it leads to the same optimal passing scores in a more circumstantial way.

A radical departure from decision theoretic methodology as hitherto used in domain referenced testing (Davis, Hickman and Novick 1977, Petersen 1976, Huynh 1976, 1977, Van der Linden (1980) is the modelling of expected remediation effects. It may seem obvious that in decisions to give remediation or not, explicit attention should be given to the expected effects of this special treatment, analogous to the role expected benefits play in personnel testing and testing for selective admission in education. In the specialized field of domain referenced testing, however, the idea that some kind of predictive validity might be involved, and that consequently validation studies are a necessity, has not been uttered (but see Wilbrink 1980), let alone worked out in a proper methodology.[26] Instead, the decision maker had to subsume his subjective notions of possible treatment effects in his utility assignments, not a very nice thing to do, and not to be excused as 'cutting the decision tree' at an appropriate level. There is simply no merit in mixing up utilities and probabilities and in substituting implicit judgments for explicit validation study results.

This new methodology, and the more careful way that utilities are assigned, will in most cases result in terminal utility functions that do *not intersect*. As in Aptitude Treatment Interaction methodology (Cronbach and Snow 1977, e.g. their figure 2.6), such intersection, or even ordinal intersection, is not necessary for a non-trivial optimal passing score to exist. This is an immediate consequence of using two probability models as weighting functions, instead of the one $f(\pi \mid x)$ as used by, for example, Davis, Hickman and Novick (1973). These authors used the intersection of the terminal utility functions to derive the optimal passing score, a procedure that implicitly assumes $u_t(r, \pi)$ to be consistent with the following equality:

$$E_{\pi \mid x}\ u_t(r, \pi) = E_{\tau \mid x}\ u_t(r, \tau) \qquad [47]$$

for all X. For subjective judgment to be this consistent would be a rather marvellous accomplishment; this consistency condition may not even have a mathematical solution. Of course, the same implicit assumption is made by authors using normal form analysis with only one probability model.

A plot of expected terminal utilities for all X = x will show an intersection, corresponding to the optimal passing score q (disregarding the discreteness of X), unless the maximum score is the optimal passing score.

An important consequence of the introduction of a probability model on domain scores after remediation is the transparant way in which utilities and costs (negative utilities) can now be assigned, and combined into terminal utility functions. This tranparence extends to the interpretation of the results of the decision analysis. It may have struck the reader as an inconsistency that costs of remediation play an important role in this decision making, while costs involved in 'regular' instruction have not even been mentioned. The reason is that at the time of decision making the regular costs or 'investments' may be regarded as sunk costs, and will in any case not count in the comparison of the desirability of *further* treatment. This being the case, there is however no objection to specifying a combined utility function on regular instructional costs and domain scores reached, although doing so may introduce complications that harm the transparency of the decision analysis. See Keeney and Raiffa (1976) on the technique involved in reducing a two-dimensional utility function to the unidimensional one that is demanded by the decision analysis. When regular instructional costs are in this way introduced in the analysis, then the costs of remediation will have to be carefully specified as only costs on top of the costs that would have been made in regular instruction to reach the domain score now being reached only after remediation, to avoid double counting of costs.

I have used the beta-binomial model, or the binomial error model as Lord and Novick (1968) have it. For psychometric purposes this model is, strictly viewed, only applicable when every examinee has his test randomly sampled from the domain. Otherwise the compound binomial error model should be used. However, for many purposes the binomial error model will be a good approximation. In regard to an application as in domain referenced passing scores, I see no reasons why the beta-binomial model should not be useful when in fact a group test is used, except for the influence that accidental difficulty of the test may have on the resulting passing score. The proper thing to do, therefore, seems to be to use more than one test form, but not as many as there are examinees, each form randomly chosen from the domain, and to use more different forms the shorter the test is.

However, this suggestion applies only to the validation study: in regular practice one may use group tests again because testing will be rather frequent and every time a new random sample from the domain will be used, in this way approaching in the long run the mathematical assumption that tests are randomly sampled for the individual examinee.

The methodology presented will result in 'optimal' passing scores, 'optimal' referring to the utility specifications of the decision maker. It should be pointed out, however, that other characteristics of the instruction and assessment situation may affect the 'optimality' of results. For example: the number of items in the test, the possibility that a change in passing score will influence (feed back to) study behavior. The decision maker, aware of those contingencies, can easily incorporate them in his strategy. Another feature of domain referenced testing is that in a particular course a number of these assessments will be made, so it might be possible to design a method where passing scores on all tests in this course are optimized simultaneously. That might be done by taking end-of-course attainments and time spent as goal variables to be optimized, and experimentally varying the passing scores.

References

Abramowitz, M, & Stegun, I.A. (Eds.) *Handbook of mathematical functions.* National Bureau of Standards, Applied Mathematics Series no. 55,Washington, D.C.: U.S. Government Printing **Office.** New York: Dover, *1965.*

Becker, S.W., & Siegel, S. Utility and level of aspiration. *American Journal of Psychology,* 1962, *75*, 115-120.

Cronbach, L.J., & Snow, R.E. *Aptitudes and instructional methods. A handbook for research on interactions.* New York: Halsted, *1977.*

Davis, C.E., Hickman, J., & Novick, M.R. A primer on decision analysis for individually prescribed instruction. Iowa City: The Research and Development Division, The American College Testing Program, 1973.

Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, *41*, 65-78.

Huynh, H. Two simple classes of mastery scores based on the betabinomial model. *Psychometrika,* 1977, *42,* 601-608.

Glass, G.V. Standards and criteria. *Journal of Educational Measurement,* 1978, *15*, 237-261.

Isaacs, G.L., & Novick, M.R. Computer-assisted data analysis - 1978, Manual for the Computer-Assisted Data Analysis (CADA) Monitor. The University of Iowa, 1978.

Johnson, N.L., & Kotz, *S. Distributions in statistics: discrete distributions*.Salt La ke City: Wiley, 1969

Keeney, R.L., & Raiffa, H. *Decisions with multiple objectives: preferences and value tradeoffs.* London: Wiley, 1976.

Lindley, D. *Making decisions*.-*New York:* Wiley, 1971.

Lindley, D.V. A class of utility functions. *The Annals of Statistics*, 1976, *4*, 1-10.

Lord, F.M, & Novick, M.R. Statistical theories for mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Novick, M.R., & Jackson, P.H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

Novick, M.R., & Lindley, D.V. The use of more realistic 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.

Raiffa, H. , & Schlaifer, R. *Applied statistical decision theory*. Cambridge, Mass.: The M.I.T. Press, 1968 (1961).

Siegel, S. Level of aspiration and decision making. *Psychological Review*, 1957, *64*, 253-262.

Simon, H.A. Statistical tests as a basis for 'yes-no' choices. *Journal of the American Statistical Association*, 1945, *40,* 80- 84. Reprinted in H. A..Simon *Models of discovery and other topics in the methods of science.* Boston: Reidel, *1977.*

Van der Linden, W.J. Decision models for mastery testing. *Applied Psychological Measurement,* 1980, *4,* (to appear)

Wilbrink, B. Some radical answers to the criterion-referenced cutting score problem (in dutch, english abstract). *Tijdschrift voor Onderwijsresearch*, 1980, *5*, 112-125.

Related publications are:

Wilbrink, B. Criterion-referenced cutting scores are easily located. *Tijdschrift voor Onderwijsresearch, 1980, 5,* 49-62. (in dutch).

Abstract

In criterion referenced testing the problem of locating cutting scores that are in some sense optimal is usually solved after intricate mathematical reasoning. The same solution is shown to be obtainable by simple arithmetics, though admittedly still based on the same debateable premises as more sophisticated approacheslike Huynh's (1976). Moreover, statistical modelling is nice in theoretical work, but not a necessary condition for sensible applications. The practitioner needs only 1) the value of a selection parameter, derived from the assessed utilities (losses) on possible decision outcomes, and 2) the scatter diagram of scores on test and referral task (or a parallel test) from students not given differential treatment on the basis of these testscores. Statistical models are useful insofar as improved estimates on the probability of success on a referral task, given testscore, are obtained.

Wilbrink, B. Some radical answers to the criterion-referneced cutting score problem. *Tijdschrift voor Onderwiisresearch,* 1980, 5, 112-125. (in dutch).

Abstract

A serious defect in decision analytic approaches to the cutting score problem hitherto has been that remedial treatment effects are rather implicitly subsumed in utility assignments to an incomplete set of possible outcomes. The correct decision analysis is presented, using results from a second validation study on students assigned to the remedial treatment irrespective their testscores. The specification of utilities, not longer being mixed up with probabilistic contingencies resulting from remediation, is now possible in a rather clear-cut way. The obvious relation to Cronbach and Snow's (1977) ATI methodology is summarily pointed out. The given approach, however correct, is still rather impractical; the best and simplest procedure is suggested to be the simultaneous optimization of the cutting scores on the set of tests, experiment-wise varying cutting scores and observing resulting achievements and needed studytimes.

Universiteit van Amsterdam
Centrum voor Onderzoek van het Wetenschappelijk Onderwijs
COWO
Oude Turfmarkt 149 /Telefoon 525 2835
10 12 GC Amsterdam

werkbespreking vrijdag 30 mei 180

AMSTERDAM,

Beste COWO-ers,

hierbij het (eerste en voorlopige)-concept van het vrijdag te bespreken artikel. Een goede titel heb ik nog niet verzonnen, misschien wordt het zoiets als 'Some methodological contributions to decision theretic standard setting', maar dan wat minder nietszeggend.

Wat nog mist: het abstract, en de discussion. Beide komen later.

Er mankeert hier en daar nogal wat aan het engels (suggesties en signaleringen zijn welkom), en aan de gekozen terminologie (daar ga ik deze week nog met de vlooienkam doorheen, ik wijs vast op het halverweg e overgaan van outcome utility op terminal utility, de laatste bewoording is wat beter in overeenstemming met de bestaande literatuur).

Het is mijn bedoeling een herziene versie onmiddellijk na mijn vacantie in te dienen bij Applied Psychological Measurement (dat tijdschrift komt met een themanr. op mijn onderwerp uit, eind dit jaar, waarvoor ik te laat ben, maar dat mag niet hinderen).

Van de 'tarditionele' besliskundige aanpak heeft Wim van der Linden een uitstekend overzicht geschreven (vdL 1980), waarvan een kopie bij mij beschikbaar is. Lezen van mijn 2e TOR artikel is overigens een aardige introductie op de inhoud van wat nu voor je ligt.

Ben-

[1] '**terminal utility function**': Ik geef dat geloof ik nergens zo aan, maar deze terminologie heb ik waarschijnlijk ontleend aan Raiffa en Schlaifer.

[2] Dit beperkt de beoordelingsproblematiek onnodig tot het zetten van grensscores. Het inzicht dat dit een niet noodzakelijke inperking is heb ik later verworven, daarin speelt juist de objectieve bepaling van nutsfuncties in het Algemene ToetsModel een belangrijke rol, rond en na 1995. De onjectieve nutsfucnties in mijn individuele model zijn in algemene vorm immers deels compensatorisch. De inperking die geaccepteerde institutuonele modelen zichzelf opleggen gaat evenwel nog veel verder: de docent moet zich maar vast zien te klampen aan dat kleine beetje reductie van onzekerheden dat hara kleine toets biedt, terwijl het natuurlijk zo is dat studenten die met een bereorde studiestrategie voor haar toets toevallig slagen, dat kunstje voor andere toetsen in de reeks van het examen niet kunen herhalen.

[3] In essentie dit is alleen maar een technisch punt. De hele techniek zou verwezen kunnen worden naar een bijlage of naar een opvraagbaar achtergrondstuk. Het vervelende voor de onderzoekers van het vigerende model is dat ze dat niet hebben gedaan, en zich hebben laten meeslepen door de techniek in plaats van de kenmerken van de beoordelingssituaties op de voorgrond te zetten. Voor het onderhavige rapport zou ik nu geneigd zijn om vrijwel alle formules eruit te slopen. Ik denk er dus ook over om niet moeizaam voor deze gedigitaliseerde versie van het rapport al die formules te reconstrueren.

[4] Dit is een enorme gemiste kans: het is kortzichtig om de aandacht op de domain score te richten. Het strategische gedrag van de student, daar gaat het om. Spits dat toe op wat in de gegeven situatie waarin de individuele student zich bevindt voor haar de optimale strategie is, gebruik daarvoor het Algemene ToetsModel zoals ik dat tot 2002 heb doorontwikkeld, of voor mijn part het tentamenmodel van Van Naerssen in de vorm die het in 1970 heeft. Maar goed, de vraag die ik aan het eind van deze alinea opwerp zet de zaak misschien toch weer in het goede spoor?

[5] Deze modelopbouw vooronderstelt stilzwijgend dat de toetsscores al binnen zijn. Een heel ander uitgangspunt wordt verkregen wanneer de beslissing geurime tijd voorafgaand aan de toetsafname wordt genomen: dan gaat het om de ruwe scores die worden verkregen, dan kan de 'state of nature' de ruwe score zijn. Wie er vervolgens nog behoefte aan zou hebben een vertaling naar domeinscores te maken kan zijn gang natuurlijk gaan, maar het statistisch model is niet meer die vertaling van ruwe naar ware scores, maar de vertaling van tevoren beschikbare informatie naar een voorspellende toetsscoreverdeling.

[6] Het punt dat in deze alinea wordt gemaakt is op zich helder en goed, al moet ik het directer formuleren dan in 1980 gedaan: als de domain score de doelvariabele is, en daarover een nutsfunctie wordt gespecificeerd, dan is bij afwijzen nog steeds diezelfde doelvariabele en zijn bijbehorende nutsfucntie aan de orde, maar dan opschuivend naar het effect van de remediatie die afwijzen impliceert.

[7] De dreigende oneindige reeks van herkansingen, door Van Naerssen in 1970 nog opgevat als een mooi aanknopingspunt om een wiskundig model op te bouwen, is in een model niet aanvaardbaar. De reeks kan kunstmatig worden afgekapt, maar dat is niet elegant en staat misschien los van de onderwijspraktijk. Het probleem verdwijnt wanneer niet de domeinscore als doelvariabele wordt geberuikt, maar de investering van studietijd door de student. In het daarvoor benodigde besliskundige model voor de student is het mogelijk de optimalisering adequaat uit te werken, zodat er inderdaad altijd een optimale investering van tijd is voor iedere situatie waarin de individuele student zich bevindt.

[8] Pro memorie: als domeinscores voor de docent - decision maker niet de gepaste doelvariabele zijn, maar studietijd van de student wel, dan gaat het dus niet om nut over domeinscores, maar om nut over studietijd. Interessant is dat nut over studietijd op een bepaald punt zal ombuigen en dalen, er is dan een volkomen ander type nutsfunctie aan de orde dan in de literatuur van het vigerende emodel en in het onderhavige rapport het geval is.

[9] Scherper geformuleerd: investeringen moeten gescheiden worden gehouden van doelvariabelen. Dat maakt het ook mogelijk om al gemaakte kosten te onderscheiden van nog te maken kosten: alleen de laatste 'tellen', gemaakte kosten mogen beslissingen niet beïnvloeden, die zijn afgeschreven op het moment dat ze zijn gemaakt.

[10] Quod non. Er is geen goede analyse van actor en doelvariabele gemaakt, dus ook niet van de rest. Bijvoorbeeld ontbreekt het ondrscheid tussen een model voor een docent en een enkele student, danwel een docent en een groep studenten, danwel een model met een reeks van toetsen die samen worden geregeerd door een overgangs- of examenregeling of nog weer een andere formele regeling (bijvoorbeeld eentje die de deeltoetsen binnen een vak regelt).

[11] Het stoeien met wiskundige fucnties als nutsfuncties lijkt weinig zinvol, maar kijk er zo tegenaan: de oefeningen met deze nutsfucnties laten de techniek zien van de oplossing, gegeven de nutsfunctie. De techniek sluit niet uit dat in plaats van een handige wiskundige functie, een objectieve nutsfucntie wordt gebruikt, 'objectief' in de zin van direct afgeleid uit de geldende examenregeling. De directe tegenwerping is dan: de oplossingen zoals die in de literatuur en in dit rapport worden gepresenteerd zijn in hige mate afhankelijk van de specifieke functie die is gekozen, dus hoe kun je dan met meer waarheidsgetrouwe nutsfuncties toch deze technieken toepassen? Ik vermoed dat die tegenwerping in hoge mate terecht is, en het duidt op ernstige ontwerpfouten in het 'vigerende' model.

[12] Lineair nut veronderstelt natuurlijk ook heel veel: volledige compensatie. Dus deze alinea is onzin. Het op voorhand kiezen van een wiskundige functie en die maar 'proberen' als nutsfunctie is uiteraard geen aan te bevelen werkwijze, het is meer een ontkenning van de geest van besliskundige analyse.

[13] Als ruwe scores de doelvariabele vormen, dan lineair nut over de ruwe scores!

[14] Dat is onhandig geformuleerd, en dus een gemidte kans. Deze nutsfucntie reduceert namelijk tot de domeinscore zelf, voegt er niets aan toe, en hoeft dat ook niet te doen. Het gaat dus telekens niet om verwacht nut, maar om verwachte domein score. Maar ged, om didactische redenen kun je dat verwacht nut noemen, om voor te bereiden op casus met andere nutsfucnties.

[15] Het is erg kunstmatig om de nutsfunctie op deelgroepen afzonderlijk te definiëren. Het kan alleen maar tot verwarring leiden. De oplossing in dit rapport is om voor alle deelgroepen uit te gaan van dezelfde doelvariabele (domeinscores) en nutsfucntie daarover, maar daarmee loop ik vooruit op de behandeling.

[16] Jammer. Beter zou zijn geweest om te zeggen dat dat toekomstige risico al is verdisconteerd in de nutsfunctie, die immers niet voor niets lineair is gekozen.

[17] Nu zou ik niet meer over kosten, maar over investeringen spreken: immers, de beslissing moet nog worden genomen, dus remediëren zal een investering vragen. Desondanks is het beter om de kosten niet in de nutsfunctie over de doelvariabele op te nemen, dan zou je ook niet die verwarring hebben van twee nutsfuncties die eigenlijk hetzelfde zijn.

[18] Bij ruwe scores als doelvariabele: het gaat dan om de scores die de afgewezen studenten uiteindelijk behalen op de volgende toetsafname, dus niet om de scores behaald op de toets warop ze zijn afgewezen.

[19] **Cost combined with utility**: Raiffa en Schlaifer combineren kosten met het nut over de doelvariabele (zie in de tekst de volgende paragraaf). In zakelijke toepassingen is dat handig: alles wordt immers al gauw in financiële termen vertaald. Maar niet in alle situaties is het handig om kosten per se vergelijkbaar te willen maken met nut over de doelvariabele. Een uitweg zou kunnen zijn, dat moet ik nog onderzoeken, om te na te gaan hoe gevoelig de optimalisering is voor de 'hoogte' van de ksoten, maar doorgaans zal die gevoeligheid behoorlijk groot zijn, anders zouden de kosten op voorhand al buiten beschouwing kunnen worden gelaten.

[20] Dit voorbeeld is niet echt een goede gedachte. Kosten van remediatie zijn complex, zoals bijvoorbeeld onderzoek naar effecten van zittenblijven laat zien, en hangen sterk af van de actor die de

kosten verondersteld wordt te gaan maken. Het betere besliskundige model mengt de kosten niet met de nutsfucntie, maar geeft ze een afzonderlijke plaats bij de optimalisering.

[21] Voor een model met een enkele student aat het om de **aannemelijkheid**, niet om de kansverdeling. Dat introduceert meteen de vraag waarom op voorhand van groepsafnames van de toets wordt uitgegaan. Het ligt veelmeer voor de hand om het model eerst voor een enkele student te ontwikkelen, en vervolgens te analyseren op welke manier je met groesgegevens een versterking van het model zou kunnen verkrijgen.

[22] Ik doe niet echt moeite om zaken toe te lichten. Het is heerlijk simpel dat het eindnut gelijk is aan μ, maar dat geldt alleen voor lineair nut van 0 tot 1. Het resultaat zou verder uitgebuit kunnen worden dan in de literatuur gebruikelijk is, maar ja, de literatuur is vooral op formules gericht, en minder op begrijpelijk maken.

[23] In een model met maar een enkele student reduceert dit natuurlijk tot x / n. Het hele 'betrouwbaarheidsgedoe' valt eruit. Elders is natuurlijk wel van belang hoeveel vragen er in de toets zitten, want dat bepaalt de spreiding van de aannemelijkheid.

[24] Hier heb ik weer eens veel te weinig toelichting gegeven. Het gaat niet om het blote verschil in verwachte domeinscore, maar om het **nut van dat verwachte verschil**, en in het geval van lineair nut komt dat in mijn model neer op het verwachte verschil zelf.

[25] Lees na het tweede gelijkteken: n boven x. Idem na het gelijkteken in [18]. Om mogelijke compatibiliteitsproblemen in de toekomst te vermijden, heb ik ervan afgezien een formule-editor te gebruiken. Alle formules zijn uitsluitend met het standaard-lettertype en met het lettertype symbol in elkaar gezet.

[26] Dit is op zijn minst een onzorgvuldige opmerking. In het themanummer van APM in 1980 gaan diverse auteurs erop in. Ik heb dit kennelijk ooit opgeschreven met de bedoeling het nog te checken, en dat laatste niet gedaan.