Vier manuscripten, die allemaal iets met herziening van 1980 Passing scores APM te maken hebben. Ik heb ze ieder een eigen kleurtje gegeven. T.z.t doe ik nog eens een poging het allemaal in elkaar te weven en te actualiseren met latere inzichten. Of is dat misschien net zo verstandig? Voor mijn SPA-model was ik nog niet oz heel lang geleden begonnen om nutsfuncties uitvoerig te behandelen, het is beter om daarmee door te gaan, lijkt me. Dan is alles wat hironder volgt, alleen om te bewaren voor misschien ooit?

==========

[Dit is een poging om 'Passing scores' te herzien, de tekst is scherper, de formules eenvoudiger, de figuren beter. Maar ik heb destijds deze poging niet afgemaakt. De tekst is van belang omdat het nog eens de puntjes op de i zet, het geeft mijn vorderingen in 1980 denk ik goed weer.

Voor annotaties zie de file van het rapport 'Passing scores.' Het is 2022, en ik probeer van de OCR-file een op mijn website publiceerbaar bestand te maken. In rood geef ik op enkele plaatsen aan wat me niet bevalt of wat misschien beter anders kan. Ook breng ik nogal wt tekstuele en notatie-wijzigingen aan in de originele tekst die uit 1980 stamt. Het werk is nog niet af, zeg maar. Mogelijk ga ik nog afscheid nemen van het Engels, een begin van een Nederlandse tekst is overigens al eens gemaakt, en ook hieronder te vinden.]

file (OCR van manuscript)
80_aantekeningen_utiliteit_etc.rtf

The most promising approach to standard setting is the decision theoretic one (Glass 1978). However, the way this approach has been implemented until now has suffered from too heavy accentuation of the mathematics involved (Huynh 1976, 1977, Van der Linden 1980), with the exception of a rather unknown study by Davis, Hickman,

and Novick (1973). The latter authors used the extended form analysis, a mathematically equivalent but significantly less complicated competitor to the normal form analysis that has been used by the former authors and that is the usual form of analysis in statistical decision theory (Ferguson 1967, De Groot 1970). On the distinction of extensive form and normal form analysis see also Raiffa and Schlaifer (1971), Davis, Hickman, and Novick 1973. Extensive form analysis, the technique I will use in this paper, is the predominant form in (managerial) decision theory (e.g. Schlaifer 1959). Raiffa and Schlaifer (1971 p. ??) showed that both approaches lead to the same optimal decision.

Optimal decision making, and for that matter also optimal standard setting, refers itself to relevant criteria, usually called 'states of nature', about which the decision maker has some information but by no means perfect information. Usually there is one particular state of nature involved, e.g. in medical diagnostics, and that state of nature will determine the outcome of the

action taken. All you need at this point of the analysis is a probability model on the state of nature, connecting it to the information available. Now, that is a good way to approach the diagnostic problem, but is it also a good way to approach the special problem that the setting of a standard or cutting score on a criterion referenced test poses?

The very first thing to do when you are planning on a decision analysis, is to analyze your problem. Any omissions made here will lead to less than optimal results further on. Now, when analyzing the standard setting problem in educational assessments you will discover that there is no direct analogy to the medical diagnosis model, making it bad procedure to further follow the standard decision theoretic approach as it surely is applicable in the medical model. **The problem in criterion referenced testing is that one of the two 'treatments' is ill defined**. It is usually stipulated that 'rejected' testees will receive remediation, be tested again, if again their score is not up to the standard they receive

remediation, etc. This is a muddy state of affairs, and has to be properly reformulated for any kind of mathematical modeling to be applicable. Depending on the particular situation you may define the situation as an infinite series, as Van Naerssen (1976) and Wilbrink (1978) did. Another possibility is to follow the validation study approach; to always send the students on to the next instructional unit after remediation, or if you like after the retesting, like Barkmeier, Duncan, and Johnston (1978). Needless to say: you not only define it this way, but you also have to act accordingly.

Another most important fact that has been overlooked in decision theoretic work on this standard setting problem is the role of the state of nature variabele of interest here: either the underlying true score (domain score) or the external criterion of further success (on the next unit). It has not been recognized that these criteria are no 'steady' state variables, but are influenced by the remediation treatment. That is even the goal of that treatment: to meliorate mastery or success. So you will need a

double probabilistic model in the complete analysis, one model for each of the treatments or decisions.

In the decision theoretic approach you specify utilities or utility functions on the outcome variables of interest. It has not been realized that in both treatments the variable of interest is the same: level of mastery (if you choose an internal criterion) or success in the next unit (if you choose an external criteri on), **and so also the utility function on that variable**. Of course you specify at least one other criterion variable: the costs of remediation and retesting.

Of course, you specify at least one other criterion variable: the costs of remediation and retesting. You've got to combine this negative utility with the utility function on the main criterion. I will show in the following that this preliminary analysis of the standard setting problem situation makes possible a clearcut utility assignment, thereby strengthening what has by many practitioners been perceived

as the Achilles heel of this application of decision theory.

In the process of specifying your utilities you will implicitly also determine the cutting point on the true score dimension. You may, of course, independently specify this cutting point, and your utilities. Of necessity both results must be in agreement with each other, otherwise the system will be undefined: you would have to choose either your utility functions or your cutting point as point of departure for further analysis. The wise man or woman will update his or her ideas on this cutting point on the basis of his utility specifications. In particular this means that Glass (1978) is mistaken in his criticism that decision theory would need a (all too subjectively chosen) cutting point on the true score dimension, and that Van der Linden (1980) is mistakingiy believing that cutting point is needed in the decision theoretic approach he promotes. The one exception possibly is the threshold utility case, where specification of four utilities as well as specification of the threshold or cutting

score or mastery score is needed.

Linear utility: an opportunity to clarify some unattended to issues in standard setting.

[Misschien is het een idee om hier de notatie van Peterson & Novick 1976 te volgen? Laatste sectie. O wacht, zij geven drempelnut als voorbeeld. Zie ook Cronbach, in het JEM-nummer er onmiddellijk op volgend.]

A domainreferenced test is used to either pass students on to the next instructural unit, or remediate them and send them after retesting on to the next instructional unit irrespective of their scores on the second test. This construction is chosen to define the decision situation. It is also a possibility to follow in actual practice, see Barkmeier, Duncan, and Johnston 1978. Another way to model the decision situation would be the rule that on the second test the same pass/remediate decision is to be made, and again on a third test, etcetera. This model is

somewhat less tractable, but will give the same mathematical results as the model here suggested (because it is the expected cost of remediation, and the expected results of remediation, that are used in the development). explain, Ben.

The **criterion** of interest is the level of mastery of the student, or the proportion correct expected, should the student answer all questions in the domain.

It is assumed that every student gets an individual test, randomly sampled from the domain. In practice you may of course relax this assumption, and still use the results of this study, for it is not to be expected that group tests will differ greatly in their relevant results.

Because level of mastery is what the decision maker is interested in, the first thing to do is specify the decision maker's utility function on this dimension. This function will for the expository purposes of this paper be chosen to be linear. Let us choose the utility scale in the most convenient way: assign utility zero tot zero

mastery, utility one to perfect mastery, or mastery one. Figure one depicts the result. In formula:

$u_m = m$ [1]

$f(u) = \pi$. [1]

π being the proportional level of mastery; alternatively: think of π as the parameter in the binomal model for testscores.
[in het origineel heb ik u(m)=m, maar dat is verwarrend want dat gebruikt m in twee verschillende betekenissen. Nu is m gewoon mastery als doelvariabele, en π de waarde van de variabele. Ik zal deze verandering dus overal in het vervolg moeten doorvoeren. En waarom schreef ik het niet als f(u)? Het is een functie, tenslotte. Toch eens in de literatuur kijken welke varianten in notatie er in omloop zijn

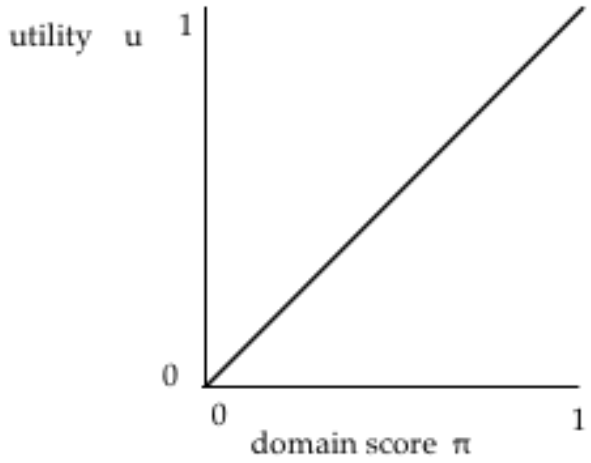


FIGURE 1. Utility function on level of mastery

If questions in the domain can be answered correctly by guessing the answer or the alternative, the same utility function may still be used. It is not necessary to evaluate the chance level of mastery on the utility scale as zero. Extending the linear function to levels below chance level need not bother you, because in applications students will always have an expected mastery not less than the chance level. The mastery dimension is in this way defined as **including** guessing. This definition will bring great benefit in the mathematical development of this model. Is this correct, Ben? Check.

Remarking that passing a student with a given test score entails the utility corresponding to his mastery level, and no other utilities or costs, you can specify the outcome or terminal function for the decision alternative 'pass' given the observed test score X:

$$u_{plm} = m \qquad [2]$$

It is identical to the utility function on level of mastery (1), depicted in figure 1.

You might object to this, feeling that a passed student having a low level of mastery will run the risk failing his next unit because of inadequate preparation. In effect you would be saying then: what I am interested in is future success, not the level of mastery. You may exchange the internal criterion, level of mastery, against an external one, let's say level of mastery on the end-of-next-unit test. In this paper I will stick to the internal criterion.

Now for the students going through remediation: here also the interest is in level of mastery reached after remediation. It is still the same mastery, defined on the same item domain, and the utility function on mastery is already specified as (1). Only, now there is a cost involved: the cost of remedial teaching, extra time spent bij the student, and the cost of retesting. For a start this cost can be assumed constant, given the observed testscore; you may think of it as an expected cost, if you wish. No other costs or utilities being involved, you can specify the **outcome utility function** for the decision alternative 'remediate' given the observed test score x:

$$U_{rlx} = m + c \qquad (3)$$

where c is the cost of remediation etc.

The cost c is to be evaluated on the already established utility scale, and you will specify it approximately as the cost of remedial teaching as a proportion of the cost the 'normal' teaching activities of this

unit entail. To be specific, let's evaluate c as .25. Both outcome utility functions can now be pictured, see figure 2. **Die keuze voor .25 is niet helemaal willekeurig, maar is natuurlijk wel onbevredigend. Is er geen betere procedure om tot een schatting te komen? In de literatuur te vinden? De behandeling van de kosten als een constante factor is een eerste benadering, waarschijnlijk goed genoeg om als steun bij het beleid te kunnen dienen.**

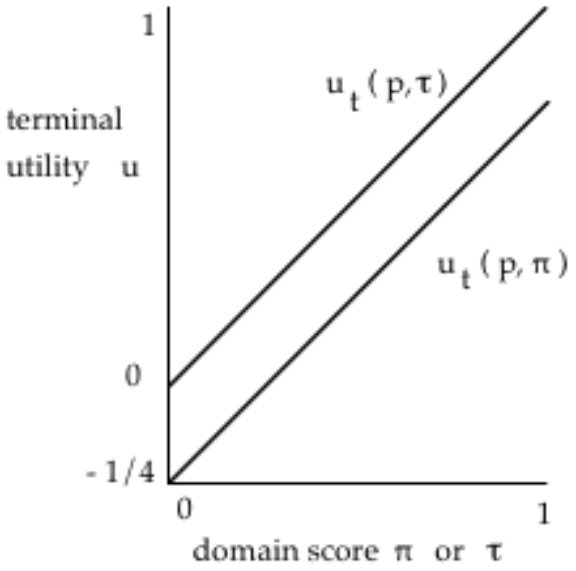


FIGURE 2 Outcome utility functions for alternative decisions 'pass' and 'remediate', given test score x. [De figuur komt uit het 1980 rapport; in de revisie heb ik horizontaal: mastery, verticaal: utiliteit; laagste punt: -0,25; onderste functie: remediation; bovenste" pass.]

This picture is remarkable, in that there is no intersection of the outcome utility functions, not even an 'ordinal interaction'. You must realize that figure 2 involves only outcome utilities, the effect of remediation is not yet taken account of. The effect of remediation is, hopefully, a 'shift' of the probability distribution over mastery in upward direction. So what you need next is the probability model on each of the decision alternatives.

23580

Considering the decision alternative 'pass', let f(M I x) be the distribution of mastery conditional on testscore x. The quantity of interest is the expected utility when the decision 'pass' is taken. To obtain this E(u

p) for every level of mastery the product of its outcome utility and its probability is taken, and the results are summed. In this continuous case the integral from zero to one is taken of m.f(MIX),

the result of which is simply the expected value of f(m I X).

Assuming the regression to be linear you may use

$$E(u_p) = E(m \mid x) = r_{xx'} x + (1 - r_{xx'}) \bar{x} \quad (4)$$

where $r_{xx'}$ is an appropriately chosen reliability coefficient for the use of this test in this situation.

Except linear regression, no distributional assumptions are involved in (4). Are you prepared to use the betabinomial model, the result is (5), as shown in the next paragraph.

$$E(u_p) = E(m \mid x) = (a+x)/(a+b+n) \quad (5)$$

where a, b, and the number of testitems n are the parameters of the beta binomial distribution BeBi(a, b, n) fitted on the observed testscore distribution.

6 23580

Considering the decision alternative 'remediate', let f(m'ly) be the distribution of mastery conditional on the testscore y on the test taken after remediation. However, the decision Iremediate' or 'pass' has to be taken conditional on testscore x, meaning that the regression of y on x, represented by g(ylx), is involved. The predicted level of mastery after remediation ml involves a double expectation, that can be simplified by taking the effect of remediation on level of mastery constant, given testscore x. You may think of this constant k as the expected gain in mastery, given testscore x.

Whether you work with k or with g(ylx), in order to implement the model in practice it will be necessary to carry out a **validation**

study on an unselected cohort that is given remedial instruction. Alternatively it may be possible to substitue some subjective estimate for the estimate based on an empirical validation study, provided it can be shown in a sensitivity analysis that the range of notunreasonable subjective estimates do not influence the standard setting.

Assuming the effect of remedial teaching on mastery constant the result for the **expected utility** when the decision 'remediate' is taken given de testscore x, and given the outcome utility function (3),:

$$E(u_r) = E(m \mid x) + k + c \quad (6)$$

Transposing f(mlx) over a distance k to the right is equivalent to a vertical shift of the outcome utility function over the same distance, because of the direction coefficiënt 1. So, take the product of outcome utility (m+c+k) and probability f(ml x) for every m, and integrate over m, the constants in (6) obtaining because the integral over a probability distribution

equals 1. Of course:

$$E(u_r) = E(m \mid x) + k + c = r_{xx'}\, x + (1 - r_{xx'})\, \bar{x} + k + c \qquad (7)$$

or $$E(u_r) = E(m \mid x) + k + c = k + c + (a + x)/(a + b + n) \qquad (8)$$

The result is only natural, as I will show here.

The **decision rule** is: select the alternative having the greater expected utility. Comparing (3) and (6) the rule stipulates: given testscore x, pass the student if c = k, in other words: **pass the student if the expected cost of remediation is greater than the expected gain in mastery** (or equal to it). You should realize that this simplicity is won by evaluating to it). the cost of remediation on the particular utility scale used.

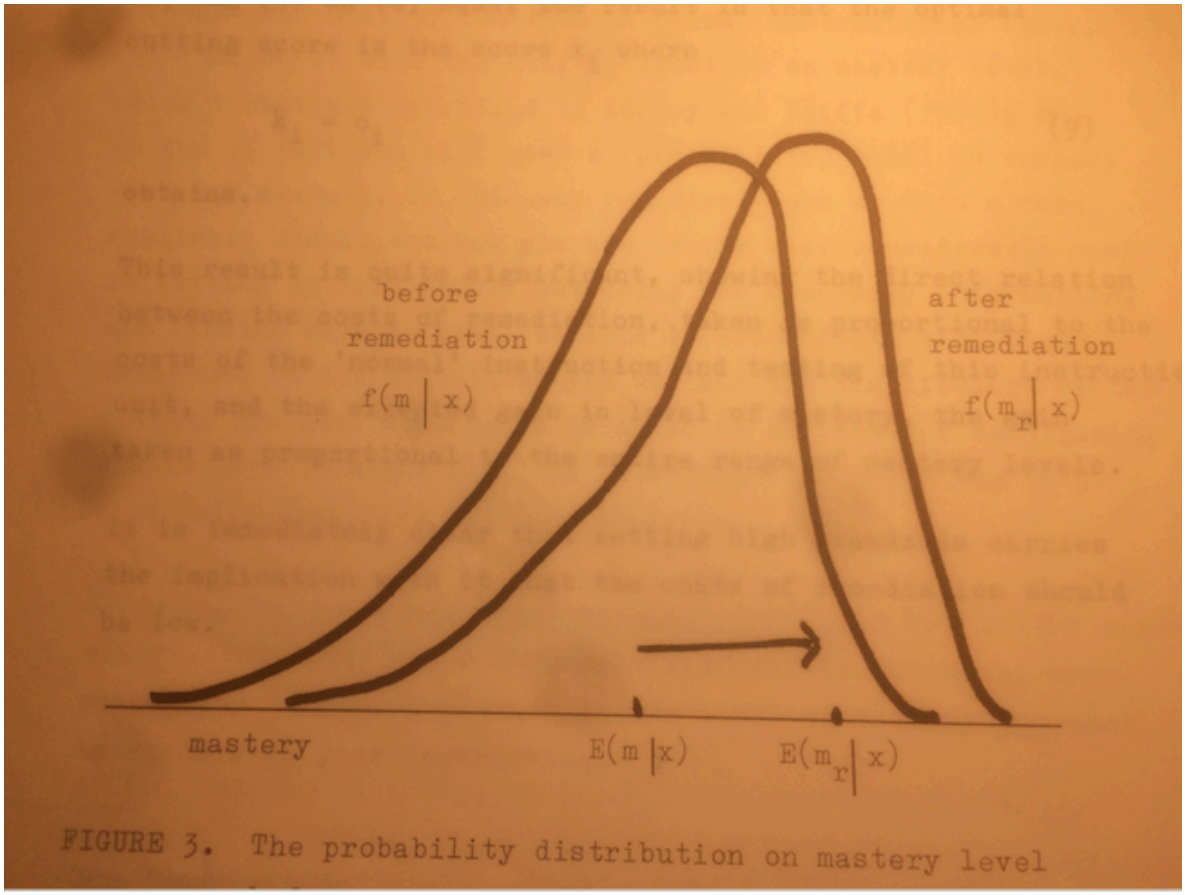


FIGURE 3. The probability distribution on mastery level

mastery E(MlX) E(mrlX)

FIGURE 3. The probability distribution on mastery level before and after remediation, given test score

8 23580

Now you may take for every observed test score x the optimal decision. In that way you will surely stumble on the test score x where for the first time you decide to pass

students having that score. Now this is your optimal standard, your best cutting score. Having a small test, stumbling will not be tedious to get at. However, it would be more elegant to be able to derive the optimal cutting score analytically.

The optimal cutting score has the unique characteristic that E(u P) and E(U .) are equal (for the moment disregarding the discrete character of test scores). It is the score where you, the decision maker, are indifferent as to the available decision alternatives. Setting (3) en (6) equal the result is that the optimal cutting score is the score x i where

$k_i = c_i$ (9)

obtains.

This result is quite significant, showing the direct relation between the costs of remediation, taken as proportional to the costs of the 'normal' instruction and testing of this instruction unit, and the expected gain in level of mastery, the gain taken as proportional to the entire range of mastery

levels.

It is immediately clear that setting high standards carries the implication with it that the costs of remediation should be low.

9 23580

When is this linear utility function on mastery level realistic? Only in the case where instruction concerns only loosely connected factual materials, and there is no interest in reaching at least some specified level of mastery regardless costs of instruction.

Using linear utility the standard on the test is set in response only to the special character of the learning curve for the instructional content concerned, and cost of extra activities in remediation and retesting, other instruction costs implicitly being taken as linear in level of mastery 'produced'. Latter remark may seem surprising, but is indeed compatible with the linear utility specification. When you are not satisfied with that particular state of

affairs, you may of course specify a (negative) utility function on normal instructional costs, and combine it with the utility function on mastery level, using techniques specified in Keeney and Raiffa (1976); and on top of that you will need a (probability) model to connect cost and mastery. In this way you might come up with a more realistic model, but now you are facing also considerable cost (opmerking in de marge: overdone) in setting up this model, gathering the necessary information, and finding ways to still be able to communicate with those concerned on the techniques you use and the significance of the results you get. Although a price is to be paid for maintaining simplicity, it might still be lower than the price tag on a more fully specified decision analysis.

However, there are ways to do better, involving but little extra effort. You might use a more flexible utility function, more responsive to your preferences in the level of mastery you want to reach with your instruction.

- -

1) The 'combination' would probably only be additive, so simplicity can be retained.

10 26580

A slightly more complicated case with linear utility.

Maybe you feel that there is some level of mastery that you prefer relatively more than other levels. More precisely: there may be a level of mastery where a small gain in mastery is preferred to the same gain at other, lower or higher, levels of mastery. You may use a rating technique to locate that preferred level, let's call it the **mastery score**, e.g. the technique used by Siegel (1957), Becker and Siegel (1962). In fact you would be deriving your utility function on mastery level, so techniques given by Keeney and Raiffa (1976) could also be used. For the purpose of this paragraph I will only use the information on the mastery score.

Now you may specify a linear utility function on mastery from zero until the mastery score d. You are free to scale it as zero at zero mastery, and one at the mastery score:

$u_m = m/d \quad \text{for} \quad 0 \leq m \leq d \qquad (10)$

Concerning higher mastery levels your preferences may be rather neutral: you regard them highly but you are aware of the maybe great outlay in extra time spent. Maybe utility could be taken as on the constant value of 1, formula(11) or even decelerating linearly, formula (12).

$u_m = 1 \quad \text{for } d \leq m \leq 1 \qquad (11)$

$u_m = d/m \quad \text{for } d \leq m \leq 1 \qquad (12)$

For the decision alternative 'pass', the combined functions (10) and (11) (or the functions (10) and (12), or some other linear variant) will also be the terminal utility function, given test score x.

Now the expected terminal utility under the decision 'pass' takes a more complicated form, because of the break at the mastery score d:

$$E(u_p \mid x) = \int_0^d (m/d)\, f(m\ x)\, dm + f\, f(m\ x) dm$$

$$\int_d^1 f(m\ x)\ dm \qquad (13)$$

For the decision alternative 'remediate' it is again assumed that the cost of remediation is the constant c, to be added to the utility function on mastery level to get the terminal utility function (c of course being a negative quantity). Expected terminal utility of the decision 'remediate' is now:

$$E(u_r \mid x) = \int_0^d (c + m/d)\, f(m_r\ x)\, dm + \int_d^1 (1 + c)\, f(m_r\ x)\, dm$$

$$= c + \int_0^d (m/d)\, f(m_r\, x)\, dm + \int_d^1 f(m_r\, x)\, dm \quad (14)$$

In this case you can't get away with it without specifying both probability models f(m x) and f(m r x). A natural thing to do when you are dealing with achievement tests is to choose the beta binomial model. The distribution of observed scores f(x) is approximated by the betabinomial distribution BeBi(a, b, n), the parameters a and b are estimated using formulas 15 and 16 (method of moments), n being the number of test items, $\bar{x}$ the mean observed testscore, s the variance of the observed testscores:

$$b = \frac{s^2 - \bar{x}(n - \bar{x})}{\bar{x} - ns^2/(n - \bar{x})} \quad (15)$$

$$a = \bar{x} / (n - \bar{x}) \quad (16)$$

Assuming linear regression of mastery on observed score, the mastery distribution

will be the beta distribution Be(a, b):

$$f(m) = B^{-1}(a, b)\, m^{a-1}(1 - m)^{b-1} \quad (17)$$

$$\text{where } B(a, b) = \frac{(a + b - 1)!}{(a - 1)!\,(b - 1)!} \quad (18)$$

12 26580

The observed score distribution given mastery will under this model be the binomial distribution:

$$f(x|m) = \binom{n}{x} m^x (1 - m)^x \quad (19)$$

What I am looking for is, however, the reversed conditional distribution f(m|x). Using a wellknown relation, it can easily be deducted that:

$$f(m|x) = f(m)f(x|m)/f(x) = m^{a+x-1}(1 - m)^{b+n-x-1} / B(a+x, b+n-x), \quad (20)$$

f(mlx) being also a beta distribution.

Now for the second probability model, $f(m_r|x)$: you might approximate this by assuming it to be also a beta distribution, having a mean value that is higher than that of f(ml x) equal to the estimated effect remediation has on the mean mastery level of testees having testscore x on the first test.

When there are indications that the variance of $f(m_r|x)$ is appreciably greater than that of f(mlx), you can also adapt the variance accordingly, using the disattenuated correlation between test scores on both tests, obtained in your validation study.

Now the mean of Be(a+x, b+nx) is

$$(a + x) / (a + b + n) \qquad (21)$$

and M being the mean effect that remediation has on students with test score x, the mean of $f(m_r|x)$ would be

$$(a+x) /(a+b+n) + M = (a+x+M(a+b+n)) /$$

(a+b+n) (22)

or the distribution Be (a+x+M(a+b+n), b+nx-M(a+b+n)) (23)

The variance of this last beta distribution is equal to:

$$\frac{(a+x+M(a+b+n))(b+n\,x\,M(a+b+n))}{(a+b+n)^2\,(a+b+n+1)} \quad (24)$$

If you correct on the basis of the disattenuated correlation r between both tests you choose (positive) constants c and d such that Be(a+xc+M(a+b+n), b+nxd? I(a+b+n)) will have the same mean as the beta distribution in formula (23), and a variance that is greater by a factor $1/r^2$. A few trials will suffice.

13 26580

To avoid repeating formulas that only in the algebraic form seem formidable, I will call

$f(m_r|x)$ the Beta distribution with parameters a'+x, and b'+n-x: Be(a'+x, b'+n-x).

In formulas (13) and (14) for the expected terminal utilities we met an integral that can now be evaluated;

$$\int_0^d m\, f(m|x)\, dm = \int_0^d \frac{m(a+x-1)!(b+n-x-1)!}{(a+b+n-1)!}\, m^{a+x-1}(1-m)^{b+n-x-1}\, dm$$

$$= \int_0^d \frac{(a+b+n)}{(a+x)}\,\frac{(a+x)!(b+n-x-1)!}{(a+b+n)!}\, m^{a+x}(1-m)^{b+n-x-1}\, dm \qquad (25)$$

where the first factor, being a constant, can be brought before the integral, and the other three factors constitute a beta distribution with parameters (a+x+1) and (b+nx).

Writing formula (25) in the form:

$$(a+b+n) \int_0^d$$

$$\frac{}{(a+x)} \int_0 Be\,(a+x+1,\ b+n-x)\ dm, \qquad (26)$$

the expected terminal utilities can now be written:

$$E(u\ p\ l\ x) = (a+b+n) / (d(a+x)) \text{ Ú}0d\ Be(a+x+1, b+nx)\ dm + \text{Ú}d1\ Be(a+x,\ b+nx)\ dm \qquad (27)$$

$$E(u_{p_1}|x) = \frac{a+b+n}{d(a+x)} \int_0^d Be(a+x+1,b+n-x)\ dm \Big/ \int_d^1 Be(a+x,b+n-x)dm \qquad (27)$$

and

$$E(u_r|x) = c + \frac{a'+b'+n}{d(a'+x)} \int_0^d Be(a'+x+1,b'+n-x)\ dm \Big/ \int_d^1 Be(a'+x,b'+n-x)dm. \qquad (28)$$

The integrals can be solved analytically with great effort, numerically, from K. Pearson's Tables of the incomplete BetaFunction, Biometrika, London 1934, or from tables of

the cumulative binomial function G(plr, n) by using the relations:

F(dla, b) = G(ald, a+b-1) x ≤ 0.5

G(dla, b) = G(bl1-d, a+b-1) x ≥ 0 . 5 (29)

for values of a and b at least equal to 1 (Raiffa and Schlaifer 1962, p. 217).

14 26580

Having evaluated both expected terminal utilities, you are ready to choose the optimal action, that being the action having the greater expected terminal utility.

Again it is clear that the evaluated effect of remediation competes with the cost c, the greater term winning the contest.

Now there does not seem to be an easy formula to derive the optimal standard analytically. You may assume monotonicity, and find the optimal standard in three or four

One remark concerning the betabinomial fit must be made. If the number of testees is small, the estimators may not be very stable, and some care must be exercised in interpreting results. However, a small number of testees or not, the decision maker will have to make his decision, operating on the best information he has, be it ever so scanty. So the model may be used even with small numbers: using the model will give you greater expected benefit than not using it. This principle of decision analysis (Raiffa and Keeney 1976, Lindley 1976 was already clearly formulated bij Simon (1943, 1977).

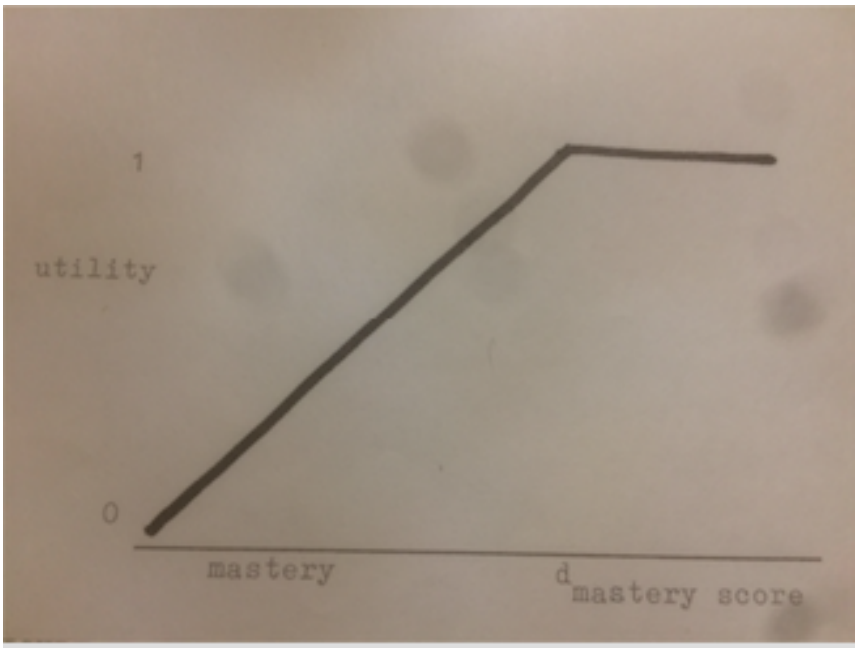


FIGURE 4. Variation, using linear utility And a most preferred level of mastery.

15 26580

Quadratic terminal utility.

A natural next step to take is to consider quadratic terminal utility. Again the decision maker determines his or her preferred level of mastery, the mastery score. Of course levels of mastery higher than this score are evaluated higher, but considering the extra instructional cost and student time involved in reaching these higher levels of mastery, it may be felt that terminal utility considering all costs and benefits is getting progressively lower once the mastery scored has been left behind by the student. Also considering there will not be a sharp distinction in terminal utility of mastery levels in the direct neighborhood of the mastery score, a smooth function like a quadratic one might be very useful. Of course, terminal utility for the decision alternative 'remediate' will again take

exactly the same form, only now the function being lower by the constant -c representing the expected cost for remediation, given test score x. Figure 5 pictures the general form of this quadratic function.

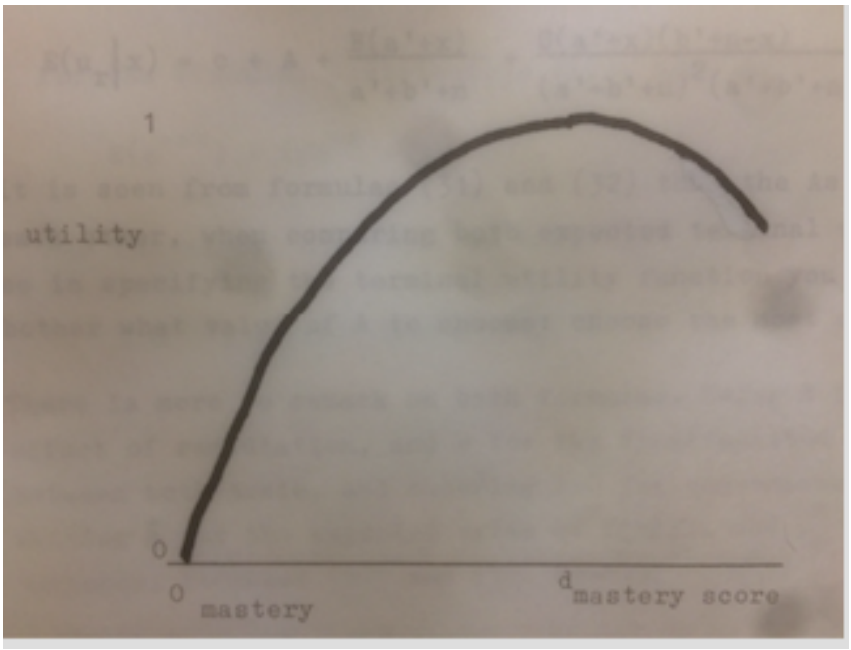


FIGURE 5. Quadratic terminal utility given test score x, for the decision alternative 'pass'.

16 26580

The general algebraic form of this function is

$$A + Bx + Cx^2$$

(30)

and you should exercise some care in choosing the specific form representing best your preferences and information.

Now this terminal utility function has the nice property that it will result in expected terminal utilities consisting of the constant A, BE(mlx), and CE((m E(m l x)^2 lx).

Using the beta binomial model developed in the last paragraph you will be able to write immediately:

$$E(u_p \mid x) = \int_0^1 (A + Bx + Cx^2)\, Be(a+x, b+n-x)\, dm$$

$$= A + \frac{B\,(a+x)}{a+b+n} + \frac{C(a+x)(b+n-x)}{(a+b+n)^2\,(a+b+n-1)} \quad (31)$$

$$E(u_r|x) = c + A + \frac{B\,(a'+x)}{a'+b'+n} + \frac{C(a'+x)(b'+n-x)}{(a'+b'+n)^2(a'+b'+n-1)} \quad (32)$$

$(a+b+n)-(a+b+n-1)$

It is seen from formulas (31) and (32) that the As cancel each other, when comparing both expected terminal utilities, so in specifying the terminal utility function you need not bother what value of A to choose: choose the most convenient one.

There is more to remark on both formulas. Using M for the mean effect of remediation, and r for the disattenuated correlation between both tests, and choosing A=0 for convenience,

writing $\bar{m}$ for the expected value of f(mlx), and s_m^2 for its variance, formulas (31) and (32) become:

$$E(u_p|x) = B\bar{m} + C\, s_m^2 \quad (33)$$

and

$$E(u_r|x) = c + B(\bar{m} + M) + C\, s_m^2 / r^2 \quad (34)$$

For the optimal standard or cutting score one obtains:

$$C\, s_m^2\, (r^2 - 1)/r^2 = c + BM, \quad (35)$$

17 26580

among other things emphasizing the importance of a greater variance that might obtain for the distribution of mastery after remediation.

Exponential utility.

When dealing with utility functions involving exponential terms , expected terminal utility can be calculated using the exponential transform T x (s):

$$T_x(s) \equiv E(e^{-sx}) = \int_{-\infty}^{\infty} e^{-sx} f(x)\, dx. \quad (36)$$

Keeney & Raiffa (1976, par. 4.9.6) give a list of these transforms for some common

probability distributions. For the beta distribution Be(a, b):

$$E\left(e^{-sx}\right) = 1 + \sum_{n=1}^{\infty} \frac{(-s)^n (a+n)!\,(a+b-1)!}{(a-1)!\,n!\,(a+b+n)!} \qquad (37)$$

For the binomial distribution Bi(p, n):

$$E(e^{-sx}) = (pe^{-s} + 1 - p)^n. \qquad (38)$$

For the normal distribution $N(\bar{x}, \sigma^n)$:

$$E(e^{-sx}) = \exp(-s\bar{x} + s^2\sigma^2). \qquad (39)$$

Special cases of exponential utility functions are the cumulative normal distribution (Novick & Lindley 1978, Van der Linden 1980), and the logistic function (Lord and Novick 1968) that might also be used as a utility function.

These exponential functions give one great latitude in specifying one's utilities, the possibilities being even greater when sums of exponentials are considered.

discussion t.z.t. [dat tezijnertijd is meen ik nooit aangebroken]

dec. theor. standard setting
ben wilbrink'begin van herziening van Passing Scores
ben wilbrink

Bij formule 2 ben ik even helemaal op hol geslagen, in de misvatting dat ik met die formule **verwacht** nut bedoel, wat niet het geval is. Afijn, er zitten ook wel een paar interessante ideeën in, dus ik heb het maar achteraan de tekst geplakt. Hier ben ik echt helemaal in de war? Nee hoor, maar ik moet duidelijker aangeven dat (1) de nutsfunctie over de doelvariabele π is, en (2) de functie van **verwacht** nut van de beslissing p (pass) gegeven x. Het is in 1980 een fatale vergissing geweest, begin k nu (2022) te vermoeden, om over outcome or terminal utility te spreken, want de ware score / ware beheersing,

gedefinieerd op de domeinscore, is niet waarneembaar, in plaats daarvan moeten we werken met het **verwachte** nut. Uiteindelijk is er in dit model nog altijd de onzekerheid over de mastery. In het SPA-model werk ik met behaalde cijfers, omdat het in het onderwijs voordurend over die cijfers gaat. Op zich levert dat de grote verwarring op tussen nut over cijfers, respectievelijk nut over beheersing. In het AP-artikel gaat het over mastery, niet over grades. Het onbevredigende van deze notatie van (2) is dat het een flauwekulformule is omdat π onbekend is. Ik kan dus beter meteen over de functie f van verwacht nut van de 'pass' beslissing spreken:

$$f(E[u \mid (pass, x_i)]) = \int_{\pi=0}^{1} u(\pi)\, L(\pi \mid x_i)\, d\pi, \quad [2]$$

for i = 0...n

[Vergelijk hoe deze formule in mijn SPA-model zit].

[De likelihood is een betaverdeling in geval van een binomiaalmodel voor toetsscores, zoveel is duidelijk. Maar hoe neem je deze integraal, mijn wiskunde is echt uitgedoofd hoor. Waarom zit ik erover in de knoop? Omdat verwacht nut wanneer de nutsfunctie over de ruwe scores gaat, heel eenvoudig is, een somformule] [De vraag die deze beslmmering dan oproept is: kan ik nut over domeinscores op een begrijpelijke manier 'vertalen' naar nut over toetsscores? Dan is de evaluatie snel opgelost. Zou het erg slordig zijn om nut over domeinscores gelijk te stellen aan nut over toetsscores, omdat de toepassing over **groepen** leerlingen gaat? Zelfs bij de cesuur gaat het al gauw om meerdere leerlingen, maar op dat specifieke punt zou zo'n ruwe gelijkstelling een bias op kunnen leveren?]

$$f(E[u \mid (pass, x_i)]) = \sum_{y=0}^{n} \left(p(y_j \mid x_i) \cdot \int_{\pi=0}^{1} u(\pi)\, L(\pi \mid y_j)\, d\pi \right), \qquad [2]$$

for i, j = 0...n (n = items in toets én herkansing)

Is dit niet een beetje een onzinformule? Ik moet het wel goed in elkaar zetten. Is dit evalueerbaar? De likelihood is een betaverdeling, bijvoorbeeld. Ik ga er verder over nadenken.

Die kans $p(y_j \mid x_i)$ is opzichzelf complex, tenzij het een onmiddellijk plaatsvindende herkansing is. In dat laatste geval is er een recht-toe-recht-aan voorspellende toetsscoreverdeling. Zit er een leer- of onderwijstraject tussen, hoe ga ik daar dan mee om? Welnu, **empirisch** natuurlijk: doe onderzoek naar dat veband, dan kan dat vervolgens gebruikt worden om zak-slaag-beslissingen op te baseren (want daar is het allemaal om begonnen). Maar zo'n empirisch onderzoek vraagt er dus om **alle** leerlingen het herkansingsproject in te studren. Is dat wel ethisch te

verantwoorden, als het praktisch al mogelijk zou zijn om iedereen voor de gek te houden? Dat is geen begaan bare weg. Zijn er alternatieve mogelijkheden om toch iets over de validiteit te weten te komen? Bijvoorbeeld een sensitiviteitsanalyse in combinatie met indirecte aanwijzingen uit andere empirische data/onderzoeken?

voorspellende toetsscoreverdeling over y moet ik nog in deze formule inwerken. Tjonge. In het SPA-model gebeurt er iets anders in module 6 http://benwilbrink.nl/projecten/spa_expectation.htm want (1) er zit een leertraject tussen x_i en de voorspellende toetsscoreverdeling (in dit geval de herkansingstoets) en (2) nutsfunctie is genomen over ruwe scores van de voorspellende toetsscoreverdeling (herkansingstoets).

==========

80PassingScoresCOWO_revision.pdf

[Om te beginnen deze brief, die vermoedelijk niet is verzonden. Ik heb waarschijnlijk telefonisch contact gehad met Wim van der Linden, die als redacteur voor APM de revisie zou begeleiden. Hij vertelde mij dat mijn stuk waarschijnlijk bij hem zoek was geraakt, ivm verhuizing in Twente. Beroerd genoeg, ik had zelf het retourstuk van Weiss evenmin ontvangen, waarschijnlijk omdat het binnen het COWO zoek was geraakt voordat ik het in handen kreeg.]

manuscript 800710 [het nummer dat mijn ingediende manuscript kreeg]

With uncomfortable long delay I have completed the revision of my manuscript for APM. The reason of the long delay is the high priority of writing a book on testitem writing techniques, the manuscript of which I finished only recently.

David Weiss encouraged me to revise the paper on two points:

(1) english sytax and idiom, and

(2) my 'knowledge (or presentation) of the criterion-referenced
testing problem'.

ad (2). There are several problems here. First, my introduction of decision-theoretic terminology was much too abstract. Second, I introduced terminology that is rather different from that used by authors with

a background in **statistical** decision theory, without explicitly noting the differences. Third, the presentation of my 'new' technique never gets convincing because there is no conceptual comparison made with the technique as for example presented by you (in APM's special issue, 1980).

Accordingly, I have written a new introduction, where I present my 'new' technique in juxtaposition to the 'old' technique. The presentation problem I have tried to solve by choosing the **theme** that decision-theoretic standard-setting techniques presuppose a true cutoff score, and showing this to be a misconception. Also I emphasize the role that evaluated effects of remediation have to play in any complete decision-theoretic standard-setting technique. The technical part of the manuscript I have kept unchanged (except syntactical and minor textual changes).

ad (1) My syntax and idiom were ridiculous, I was simply not aware of the rather stringent rules of syntax of the english language. In the new manuscript I have checked syntax and idiom to the best of my knowledge (with the help of friends). Also I will have the manuscript corrected by some knowledgeable friends.

abstract

The decision theoretic approach to standard setting on (criterion-referenced) tests is still the most promising

among the empirical approaches. Nevertheless, practical application is hindered by some conceptual difficulties. The most important misconception is the idea that a true cutoff score is assumed in all desision theoretic techniques, imparting them with a high degree of what Glass (1978) calls 'arbitrariness.' It will be shown that decision theory does not prusuppose the existence of true cutoff scores. Another crucial issue is the assignment of utilities to possible decision outcomes. In the known literature, as for example reviewed by Van der Linden (1980), no applicable techniques for the determination of utilities are presented. One of the reasons for this omission may be found in the neglect of the role that expected results of remedial instruction given to failed examinees must necessarily play in the assignment of utilities. A careful analysis is presented of the utility structure of the decision situation, leading to a transparent decision-theoretic technique for standard-setting in a very straightforward manner. Mathematical techniques are presented that enable the use of utility functions in linear, quadratic, exponential and normal distribution form, as well as of threshold utility. Numerical examples are given of the application of the technique, for several of the mentioned utility function possibilities.

Terminal utility and utility proper

In the typical presentation of decision-theoretic models

the utilities involved are to be taken as the **terminal** utilities, i.e. the utilities entailed when corresponding decisions are or would have been taken. In pass-fail decision making there are two possible actions to be taken, passing or failing xx the examinee, so there are two terminal utility functions, each contingent on one particular action or decision. In statistical decision theory it is not usual to further investigate the decision problem in terms of utilities proper, because terminal utility functions are simply regarded as given. what exactly these terminal utility functions represent remains somewhat obscure. What is worse, however, is the suggestion emanating from this approach that the decision maker somehow or other will have to express his preferences directly in terms of terminal utilities, i.e. in terms of utilities 'truly' resulting from the possible actions.

Only in special cases is it possible to directly assess terminal utilities, in most cases it will be necassary to first assess one's utilities over the possibly quite different criterion variables that are relevant tot the optimization of the decision procedure. In the digression on threshold utility an example has already and terminal utility. Here I will explain the difference in the context of linear utility.

The linear utility function on domain scores has been determined as u(d) = pi. First consider the decision to pass a student with a given observed testscore x: when no other (cost or) criterion variables are relevant, the terminal utility entailed by this decision is equal to the utility of the true score of this student.

In this case the terminal utility function $u_t(p,pi)$ contingent on the pass decision is equal to the utility function on the domain scores.

Now consider the decision to fail the student: how do we get at the terminal utility function $u_t(r,\tau)$? The most important criterion variable will still be the domain score of the student, and the utility function on domain scores of course is u(d) = pi. But what domain score are we talking of exactly? Failed students are provided with one or another form of remedial instruction, hopefully resulting in a higher domain score (better mastery). Well, then, the criterion variable relevant to the decision to retain students is the domain score reached been given of this difference between utility proper ..... (blz. zoek of weggedaan?)

Figure 2b gives the Generally Accepted Picture's (GAP) typical assumption: two terminal utility functions, defined on the same true score (domain score) scale underlying the test used, and crossing each other (otherwise a non-trivial cut-off score would not exist).

The problem with the Generally Accepted Picture (GAP) is, again, that it is left unclear how the terminal utility function for the fail decision is to be determined. The difficult task of the decision maker is to

(1) determine what will count as criterion variables for the results of remedial instruction,

(2) intuitively and implicitly determine the appropriate utility functions in accordance with the utility scale that is implicit in the terminal utility function of the pass-decision,
(3) use magic to translate this implicit terminal utility function to the domain score scale that $u_t(p,pi)$ is placed on, and
(4) use more magic to correct somehow or other for the expected effect of remediation conditional on true scores pi.

No wizards or supercomputers will be able to do this properly. A proponent of GAP may counter with the observation that the decision maker need only evaluate the seriousness of decision errors, for surely that is what utilities in the last resort are? That is nonsense, however, because decision errors can only be spoken of in the case threshold utility is appropriate and that is not the case under scrutiny. Seriousness of 'errors' must one way or other refer to expected positive or negative results of remediation or foregone remediation, and that brings us back to where we started. Wilbrink (1980a and b) gave a great deal of attention to this impossible predicament GAP leads the decision maker in, also when threshold utilities are appropriate.

Is it not possible to 'reconstruct' GAP's $u_t(r,pi)$ using the results of the complete decision analysis based on terminal utility functions like those in figure 2a? I don't think so, for that involves a method of projecting back, the mathematics of which are unclear: (1) determine expected terminal utility function on observed scores (next paragraph); 2)

translate these (how?) to pseudo terminal utilities on the underlying true score scale, and (3) because this will not result in a mathematical function, some kind of datareduction (how?) will have to be done. Improbable.

4

Now this is really quite serious, it means that GAP does not allow a realistic determination of its terminal utility function $u_t(r,pi)$, unless it could be shown from which utility functions proper it is the combination. As far as costs of remediation are concerned, it can be maintained that, when these cost may be taken to be constant for every given true score, these costs are represented in $u_t(r,pi)$ in the form of a vertical translation equivalent to the evaluated cost. But how to explain the inclination of u t(r,pi),and its elevation corrected for the utility of costs? I have no answer.

The assumption that a 'true cutoff score' is given

It was Glass who directed attention to the declared assumption of decision theoretic standardsetting methods that somehow or other a true cutoff score already should be available. (Glass, 1978). In all cases as cited by Glass, this true cutoff score was the demarcation of 'mastery' versus 'nonmastery'. As far as these methods used threshold utility functions,

this is not to be wondered at, because the choice of threshold utility implies or is equivalent to the choice of a true cutoff score. I will return to this particular case later. The remarkable point in Glass' remark was that also the use of linear utility functions by Van der Linden and Mellenbergh (1977) seemed to depend on this true cutoff score.

"*But in their application of the technique, they accepted as the criterion of mastery the instructor's opinion that 80 % of the items correct was indicative of 'mastery'.*" (Glass, 1978).

And indeed, this assumption was explicitly made by Van der Linden and Mellenbergh (1977), as it also was by Van der Linden (1980), also cited below). Now the review by Glass was fairly thorough, nevertheless he missed one important report, by Davis, Hickman and Novick (1973). These authors present a (Bayesian) decision theoretic standardsetting method that does not assume the existence of a true cutoff score, except in the case of threshold utility. In using linear, exponential, etc. utility functions a 'true cutoff' is not assumed, need not be determined, plays no role whatsoever. The question is, then: was the true cutoff score assumption of Van der Linden and Mellenbergh (1977) necessary? It was not, as I will show later. The critical review of Glass (1977) was highly influential, and left its traces in the later reviews by Shepard (1980) [shepard heeft een belangrjke noot geplaatst bij het artikel van Glass, die het niet heeft

over standards voor deeltoetsjes, maar voor national assessment], Van der Linden (1980) and Traub and Rowley (1980). Shepard (1980) on empirical methods for adjusting standards:

"*They include the more technical methods and convey the impression that standards can be determined with scientific precision. In truth, however, these approaches do not determine 4 standard; rather, they presume that a standard already exists on an external criterion and merely translate this into a cutoff score on the test.*" Shepard, 1980, p. 459.

The verdict of Shepard is strongly underscored by Van der Linden in his review of decisionthereoretic models:

"*(...) it is emphasized that in what follows two different cutoff scores are involved the true and the observed cutoff scores. Decision theory can not be used to set the former; it can be used to set the latter after a solution to the former has been obtained. (...) Though this has not always been seen (e.g. Glass, 1978) the decisiontheoretic approach to criterionreferenced testing is thus no standardsetting technique but technique to minimize the consequences of measurement Usampling error, which, preferably as a part of the normal routine, ought to follow each time a standardsetting technique is used.*" Van der Linden, 1980, p. 470.

Van der Linden can be shown to be wrong on this

important issue: the secondary role assigned to the decision-theoretic approach is unduly pessimistic. What reason could Van der Linden have to entertain this opinion? he surely gave no mathematical evidence for it. Again, it is a particular Picture of criterion-refenced measurement that plays havoc with conceptual clearness. Van der Linden formulates his Picture on the same page (1980. P. 470) as follows.

"*One of the principal uses of criterion-referenced measurement is in the assignment of students to mastery states. Typically, this involves the selection of a cutoff score on the criterion-referenced scale T. Students with true scores exceeding this cutoff score are considered masters; they are deemed to have reached the learning objectives and may proceed with the next unit or task. Students below this cutoff score are the nonmasters; usually, they are provided with extra learning time or remedial teaching.*"

This Picture is adequate as regards cases where threshold utility can be used, i.e. where there is a sharp and clear demarcation between mastery and nonmastery ('states'). But Van der Linden stretches the Picture to include cases where mastery is less an 'all-or-nothing' state, and more a gradual thing, so that linear and other utility functions are more adequate than threshold utility.

A digression on threshold utility

Utility is the qualified evaluation of (possible) results of decisions, treatments, etcetera. In surgery, for example, the most important criterion 'variable' is ultimate death or life; as long as no comparison with other criterion variables is involved, you may choose your u scale for utility freely, A good choice (from the mathematical point of view) is to assign utility zero to the possible result 'death', and utility one to the possible result 'life'. Patient and surgeon surely can agree on this. Now, for all practical purposes, 'life' is not a gradual attribute, so these are merely two different utility assignments, and no threshold utility function is involved. In criterion-referenced decision-making the important criterion variable is level of mastery, and this is a gradual attribute, to be quantified for example as the percentage of questions from a circumscribed domain the examinee would answer correctly when given the opportunity to answer all items in the domain.

In (very special) cases there may be reasons to assign utility zero to all levels of (non)mastery below a certain percentage, and utility one to all levels of mastery above this percentage or 'true cutoff score'. This is an example of a threshold utility function, the threshold being the true cutoff score, and the 'height' of the threshold depending on the difference between utilities assigned to scores to the left, respectively to the right of this true score. The choice of zero utility for non-mastery, and utility one for mastery has no particular significance except mathematical and conceptual expediency. Utilities

assigned to other criterion variables, expenditure on remedial instruction for example, -must oil course be compatible to the already set utility scale. In a particular instructional situation the cost of remedial instruction could be evaluated as for example -.2, meaning that mastery reached after remedial instruction is evaluated as having utility 1 minus .2 is .8. For technical details see the paragraph on the threshold utility function, later in this article.

[fig. 1 here appr, with threshold and linear utility]

A digression on linear utility

In criterion-referenced measurement the criterion variable usually is the proportion of items from a well-defined domain that the examinee would answer correctly when given the opportunity. Or, in the words of Harris and Stewart (1971) as cited by Glass (1978):

"*A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed.*"

Let's call this proportion of the **domain score** or **true score** pi. This domain score is the student's realization of the educational objectives (it's them

that were translated in the definition of which items belong to the domain), evaluation of domain scores or will normally result in a positively sloped utility function on domain scores. The linear function is the most simplest example, it expresses an evaluation of domain scores that is proportional to the domain scores themselves. its general form is u(d)= a + b(pi), but when you are still free to choose your scale of utility, it may be expressed as u(d)=pi. and Figure 1 presents this nice and uncomplicated function. Remark that no true cutoff score is needed to determine this utility function, the only thing the decision maker has to be sure of is that he or she evaluates (differences in) domain scores proportionally.

Utility as a function of (distance from) the true cutoff score

In the Picture that somehow or other 'masters' are to be differentiated from 'non-masters', threshold utility is the paradigm case, at least until second thoughts begin to trouble the researcher's mind. Van der Linden and Mellenbergh (1977):

"*An obvious disadvantage of the (threshold loss function) is that the loss is constant. For instance, a not-accepted examinee with a latent score just above the cutting score gives the same loss as a not-accepted examinee with a latent score far above the cutting score. This constant loss can be by using*

a linear loss function."

[Note) The term 'loss' as Van der Linden and Mellenbergh (1977) and Van der Linden (1980) use it, designates the **negative** of what usually is called terminal utility.]

This argument is contingent on the assumption of a true cutoff score, an assumption that is adequate exactly al long as the use of threshold utility is adequate. The moment threshold utility is rejected because is does not adequately capture one's evaluation of different true scores in that part of the scale of the criterion variable that roughly corresponds to 'mastery', the assumption of a true cutoff score loses its right of existence. It is not consistent to use a true cutoff score in ascertaining one's preferences over donain scores in the form of a linear utility function. But that is exactly the approach of Van der Linden and Heilenbergh (1977), because in their approach assigned utilities are proportional to the difference between the domain score pi and the true cutoff score d. This is important, because this formulation indicates the method to be followed in ascertaining one's preferences in terms of domain scores: the method of Van der Linden and Mellenbergh presupposes the true cutoff score.

In the digression on linear utility is was shown that any linear utility function may be reparameterized in the form u(d) = pi, when this reparameterization is used

to determine the scale that utilities are expressed on. The amazing conclusion is that also Van der Linden and Mellenbergh's (1977) linear utility function on domain scores may be written as u(d) = pi, i. e. **without mention of a true cutoff score**. And of course, the general form u(d) = a + b(pi - d) contains the true cutoff score d, but in no way does a linear function written as a function of a particalar score depend on that score. (it merely functions as a reference point, a new origin).

This particular coincidence in Van der Linden and Mellenbergh's (1977) treatment of linear utility triggered the hardening of the Generally Accepted Picture (GAP) that decision-theoretic standard-setting presupposes a true cutoff score.

Expected terminal utilities are decisive

Expected terminal utilities have not been mentied earlier, because I concentrated on conceptual issues concerning utilities and terminal utilities as used in decision-theoretic standard-setting. But ultimately it must be these expected terminal utilities that are decisive with regard to the decision to be taken that is optimal. The optimal decision is the decision with the higher expected terminal utility. It is terminal utility that is expected, because as you will have noticed this terminal utility is not connected to observed testscores, but to true scores. (this is not necessarily so, because a decision-theoretic

approach to standard-setting is possible with utilities proper and terminal utilities defined on obervable criterion variables; Wilbrink 1980a and b has given this technique, also in juxtaposition to GAP). Remember the digression on threshold utility: passing a student with true score below .8 results in utility zero, passing a student with true score .8 or higher results in utility one; the expexted utility of passing a student with observed score x is equal to the probability that his true score is .8 or higher.

The expected terminal utility of retaining a student for remediation involves the conditional probability density function on true scores **after** remediation: retaining a student with true score **after** remediation of .8 or higher results in utility .8, true score lower than .8 results in utility minus .2. The total expected utility of retaining a student is .8 times the probability of a true score after remediation of .8 or higher, minus .2 times the probability of a lower true score after remediation.

In this way it is possible to calculate both expected terminal utilities for every observed score x. Graphically connecting these points results in the expected terminal utility function of passong, respectively retaining examinees; where these functions cross each other, the optimal cutoff score is located. it is these expected terminal utility functions that Cronbach and Gleser (1965, 1957) called expected payoff functions, or payoff for short. For practical purposes it is sufficient to calculate algebraically the point where both expected terminal

utilities are equal to each other, this corresponds to the optimal cutoff point score. The expected terminal utility functions need not be determined, and neither the (summed) expected terminal utilities as is usual in so-called normal form analysis as preferred in GAP.

So called 'decision errors'

True cutoff scores are not presumed by decision-theoretic models. Not only that, there is no place for this concept in those situations where continuous utility functions instead of threshold utility are chosen. Where could in the case of linear utility on domain scores, a 'true cutoff' be located? However, there is an opportunity to use another concept: the preferred true score, that has a remote resemblance to the true cutoff score. I will introduce the preferred true score in the paragraph on quadratic utility. When there is no useful place left for these true cutoff scores, there obviously is no reason to artificially reintroduce them. However, this is exactly what is usual in certain kinds of consistency, reliability or validity analyses. In the words of Traub and Rowley (1980) (I could have cited many others, including Taylor and Russell, Cronbach and Gleser, or myself for that part):

"*If, corresponding to cutoff score c on the observed score scale, there is a cutoff score y on the true score scale, then the ideal or true decision would be*

to pass a person if his or her true score T equals or exceeds y; otherwise, the person should be failed. A comparison of the decision based on the observed score x with that based on the true score τ reveals whether or not a decision error has occurred."

The curious situation is that an optimal cutoff score has been determined without in any way involving the concept of a true cutoff score, and then this 'true cutoff' is calculated or projected back from this optimal cutoff score! But this 'true cutoff score' does not correspond to some meaningful state of affairs, it is a mere straw man, put up to serve the purposes of the believer in reliability analysis. Only when threshold utility is appropriate, there is also a legitimate place for the 'true cutoff score' concept, but then it is equal to the threshold chosen (two thresholds in fact, the one on the criterion variable appropriate for the pass decision, the other on the criterion variable that captures the effects of further treatment given after the fail decision). And threshold utility may only be appropriate when some kind of real classification is to bee made (man and women, has phenylketonuria or not, is a suicide risk or not), i.e. those cases where the use of (multiple) discriminant analysis is appropriate (compare Cronbach and Gleser 1965 p. 115).

The implication is that only in case of a real underlying classification it makes sense to speak of 'decision errors'. When for example linear utility is used, there is no sense in which the concept of 'decision error'

might be appropriate. There simply is no room left for the common viewpoint among psychometricians, as adequately expressed by Shepard (1980):

"*The most important point, which will influence the choice about whether or not to set standards, is that there is always error attached to the selection of cutoff scores. Individuals immediately on either side of the standard will be virtually indistinguishable from each other. With a good test, valid distinctions can be made between those who are well above or well below the standard; but passfail distinctions near the cutoff will have poor validity because a continuum of performance has been 'arbitrarily' dichotomized.*"

This 'most important point' is quite beside the point, because the implicit reference to a true cutoff score is inappropriate. That a particular examinee with a score very to the cutoff score on the test, may pass or fail under influence of luck, is quite another matter. The quality of the decision procedure may be influenced by using a better or a longer test. The index of 'quality' is the total expected terminal utility, and what keeps you off from the ideal of perfect decisions is the extra cost involved in lengthening your test. But this problem is not a standard setting problem. The psychometrician will never be able to protect examinees against bad luck, and will never be able to give satisfactory reasons to the examinee that is only one point short of the cutoff score, why he should fail. Luck and bad luck are part of the game. More formally: you juridical agreement on the

way passfail decisions will be taken.

When reference to a true cutoff score is not appropriate, when decision errors cannot properly be spoken of, it follows that also the study of consistency of decisions when using parallel tests is quite beside the point. Traub and Rowley (1980):

"*A second approach to describing the goodness of the dichotomous decision situation is to study the correspondence between decisions based on the observed scores for two or more parallel tests. This approach, too, has its analogue in traditional reliability theory, the correlation between observed scores on parallel tests.*"

This is a study in reliability of the test, not of the decisions taken.

==========

[Yet another manuscript text on the same subject, mogelijk is dit toch een latere versie dan het manuscript hierboven]

Ik heb alsnog in juni 2002 een paar wijzigingen ingevoerd uit oude aantekeningen.]

Conceptual issues in the search for cutoff scores on domainreferenced tests.

Contrary to current opinion, decision models (Van der Linden/ 1980) can be shown to optimize cutoff scores on domainreferenced tests in a nonarbitrary way (Glass, 1978), i.e. without the use of pre established standards.

posing the problem

Recent reviews, e.g. in Berk (1980a) and in the 1980 Special Issue on CriterionReferenced Testing Technology of this journal, have it as part of the 'Current Opinion' on cutoff scores in domain-referenced testing, that there no method in setting cutoff scores that does not presuppose a known 'true' standard in one way or another. The methods attributed to Nedelsky and to Angoff presuppose that the decision maker has a clear mental picture of the Minimally Competent Person. Methods that separate known groups of masters and nonmasters, e.g. the method of discriminant analysis as used by Zieky and Livingston (1977), beg the question in a most obvious way. Since the presentation of Mellenbergh and Van der Linden (1977), it is said even of decision models that "they are techniques for minimizing the consequences of measurement and sampling errors once the true cutoff has already been chosen" (Berk, 1980b), or that "the mathematical work in choosing a cutoff score along the observed score scale starts with the assumption that a standard has already been defined, either on the true scale or on the scale of the criterion measure; how this standard gets defined is never dealt with satisfactorily" (Traub and

Rowley, 1980). However, decision models can be shown to be independent of any preestablished standards. Another bias in Current Opinion is the neglect of cost and effects of the (remedial) treatment that is to be given to failed examinees. To set the stage for the extensive treatment of both themes, a description of the cutoff score problem in domain referenced testing will be given.

A *domain-referenced test* is constructed to assess the performance levels of examinees in relation to a set or domain of welldefined tasks, objectives or competencies (cf. Hambleton, 1980). In this article, the domain of tasks will be taken to be the item pool from which the testitems are sampled. This item pool actually exists, or may be taken as "a convenient conceptualization" (Wilcox, 1980). The *domain score* represents the proportion of items an examinee would answer correctly if the examinee were to answer every item in the domain. Note that the domain score is a generic true score, not the true score specific to the test used. The domain score is not corrected for guessing (see Wilmink and Nevels, 1982, on domain scores that are corrected for guessing). *Domain*-referenced is preferred to *criterion*-referenced, because the domain is what is referred to, and calling the domain the criterion invites misunderstanding (as history shows, see e.g. Hambleton 1980).

Cutoff scores on domain referenced tests will be needed in some cases, e.g. in certification (cf. Shepard, 1980) or in Individually Prescribed Instruction (IPI) to monitor the students progress

from unit to unit. It is the latter use, called pupil diagnosis by Shepard, that is focused upon in this article. That includes annual tests for gradetograde promotion, where the cutoff score should correspond to the point where the expected (evaluated) results of repetition of a grade of schooling balance the expected (evaluated) results entailed by passing the student with an observed score equal to the cutoff.

The critical comment of Jackson (1975) emphasizes this point.

"*Very seldom is there any substantial help provided to repeating pupils; instead, they are recycled through a program that was inappropriate for them the first time and that may be equally inappropriate and of less interest to them the second time.*"

'Cutoff score' is a more neutral term than 'Passing score', ladvancement score' or the equivocal 'standard'. In some cases, students may be retained for remedial instruction only after scoring below the cutoff in two or three consecutive unittests. Cutoff scores also need not refer to a 'standard' or 'true cutoff score' specified on the domain score scale.

Now the crucial question is how to choose the cutoff score on a given test, used in a given instructional situation. What the best or optimal cutoff score is will depend on (a) the educational objectives, and (b) the cost(s) of remedial instruction. To keep the problem tractable, the decision situation has to be specified carefully. In the following a simple but realistic situation will be employed, to highlight the the

conceptual issues involved. The quality and duration of regular instruction, as well as of remedial instruction, are supposed to be given. Enhancement of instructional quality is not at issue here. The cutoff score is meant to separate students who may proceed to the next instructional unit or grade from students who are to recieve remedial instruction. The domain is supposed to be specified already; and the number of items in the test is supposed to be given. Note that optimizing cutoff scores is not the same kind of problem as that of finding the optimal number of items to use in the test (see Wilcox, 1980, on the latter topic). Finally, the realized domain'score is supposed to be the only educational objective to be maximized, and the cost of remedial instruction is the only cost to be minimized.

In this decision situation the optimal cutoff score might be derived by first solving this problem: Given a student with a particular observed score, if he or she were to be given remedial instruction, would the expected gain in his or her domain score be worth the extra cost? Whether the latter specific formulation, or the former more general formulation is chosen, the solution is to be found only through decision theory.

two kinds of decision theory

Most studies on the optimal cutoff score problem have been inspired by what is called statistical decision theory (presented by e.g. Chernoff and Moses, 1959, or DeGroot, 1970), the science of decision making

under uncertainty. However, the handling of uncertainty is certainly not the only feature of statistical decision theory, as the influential statement of Van der Linden (1980) seems to suggest:

"*(..) the decision-theoretic approach to criterion-referenced testing is thus no standard-setting technique, but a technique to minimize the consequences of measurement and sampling error, which, preferably as a part of the normal routine, ought to follow each time a standard-setting technique is used.*"

The whole point in using statistical decision theory is its explicit treatment of the utilities of the expected outcomes contingent on the decision alternatives. (The utility of an outcome is a particular evaluation of that outcome). Typically, statistical decision theory takes it for granted that the decision maker has already determined the relevant utilities.

Statistical decision theory will handle the uncertainties involved in the cutoff score problem. Which uncertainties? Measurement and sampling error were already mentioned. Another uncertainty involved is the gain in domain score, that is due to remedial instruction. The decision models presented by Van der Linden (1980) do not handle the latter uncertainty, because it is implicitly absorbed in the assumed utlities. I will come back to this particular point later.

Now the only problem left is how to compare the gain in domain score to the remediation cost involved. It is

here that decision analysis, the twin brother of statistical decison theory, has its part to play. Decision analysis (e.g. Raiffa, 1968, Keeney and Raiffa, 1976, LaValle, 1977) supplies techniques to evaluate possible outcomes on a common utility scale. The utility function on the domain score scale is the concrete expression of the evaluation of different domain scores has found; the cost of remediation is to be expresssed as (negative) utility also. Bringing domain scores and remediation cost on the same scale of utility makes it possible to compare them. The solution of the optimal cutoff score problem will then be given by the techniques of statistical decision theory, using the utility structure as derived through decision analysis.

The decision situation and its approach sketched above summarize the exposition that is to follow. First the cutoff score problem will be analyzed conceptually, in juxtaposition to what I have called Current Opinion. Then the technique to find the optimal cutoff score will be developed in a constructive way.In actual practice the technique may be employed in the way it is presented here. However, the main purpose of the more technical part of this article is to conceptually clarify the cutoff score problem. The acquired insight in the cutoff score problem may lead to solutions that do not employ the technique that is presented here. For example, in Individually Prescribed Instruction optimal cutoff scores on the unittests may be found simultaneously through experimenting. Finally, an important indirect use of the technique to be

presented is in simulation studies, e.g. sensitivity analyses.

Current Opninion and the conceptual issues it entails

In this section I will discuss the conceptual issues that arise when decison models are used to optimize cutoff scores on domain referenced tests. In this discussion I will introduce and illustrate the concepts from decision theory as they are needed. The concept of *threshold utility* is central to the notion in Current Opinion that decision models would assume pre-established true cutoff scores, so this concept will first be presented.

Digression on threshold utility

In decision making, the best decision alternative is the alternative that results in the outcome that is the most highly evaluated or that has the highest utility. When outcomes are uncertain, the best decision alternative is the alternative having the highest *expected utility*, the expectation being taken over all outcomes that are possible under the decision taken. For example, in surgery the important outcomes may be 'life' or 'dead; 'life' may be assigned utility one, 'dead' may be assigned utility zero. This particular choice of utilities establishes the utility scale on which all other utilities, for example the utility of a crippled life when surgery is refused, are to be

expressed. In this illustration the expected utility of the decision to operate is equal to one times the probability of a succesfull operation; if this expected utility is higher than the utility of not operating, the rational thing to do is to assent to the plan for surgery. Because 'life' or 'dead' is not a gradual attribute, the utility assignment is discrete and not in the form of a threshold function.

In domainreferenced testing, the relevant outcome is a gradual attribute: the domain score or the proportion of items in the domain the examinee would have answered correctly were he to answer every item in the domain. In some (very special) cases there mey be reasons to assign utility zero to all domain scores below a certain proportion, and utility one to all domain scores equal to or higher than this proportion. This would be an example of the assignment of a *threshold utility function* to the domain score scale, the threshold being located at what is often called the true cutoff score or the mastery score. Utilities that have to be assigned to other outcomes or attributes will have to be expressed on the utility scale that is established by the (freely chosen) threshold utility function on the domain score scale. For example, the cost of remedial instruction might be evaluated as equal to . 2 on the established utility scale. If the cost of remediation is a constant, i.e. independent of domain scores, then the utility of a particular domain score as reached after remediation is to be dimished by .2. For example, domain scores higher than the true cutoff score are evaluated as having a *terminal utility*

of 1 .2 = .8. The decision to retain the pupil or not will have to be taken in the face of uncertainty, because the domain score is not known. In the example given, the pupil will be retained for remedial instructed only if the expected result of remedial instruction, given the observed score on the test, is higher than a difference of .2 in the probability that the domain score of this pupil is above the true cutoff score. For the technical details involved, see the paragraph on threshold utility in the next section.

The assumption of a pre-established true cutoff score

It was Glass (1978) who directed attention to the fact that authors of papers on the use of decision theory in optimizing cutoff scores expressed the explicit assumption of a pre-established true cutoff score or mastery score on the domain score scale. This assumption is quite natural when threshold utility is appropriate, because the assignment of a treshold utility function on the domain score scale implies or is equivalent to the choice of a true cutoff score or mastery score. I will later on return to this particular case. What was really remarkable in the point made by Glass, was that the use of linear utility functions by Van der Linden and Mellenbergh (1977) also seemed to depend on pre-established true cutoff scores: "*But in their application of the technique, they accepted as the criterion of mastery the instructor's opinion that 80 % of the items correct*

was indicative of 'mastery'." (Glass, 1978). Had Glass read the important report by Davis, Hickman and Novick (1973), he would certainly have corrected his judgment. These authors presented a (Bayesian) decision-theoretic technique to optimize cutoff scores, not assuming pre-established true cutoff scores. When using linear or other continuous utility functions, a true cutoff score need not be assumed nor determined, it plays no role whatsoever. The question then is: was it necessary for Van der Linden and Mellenbergh (1977) to assume a pre-established true cutoff score? It was not, as I will later on show. However, the critical review of Glass (1978) was highly influential, and especially so on this point of the assumption of pre-established true cutoff scores, and it left its traces in the later reviews by Shepard (1980), Van der Linden (1980) and Traub and Rowley (1980). I cite Shepard on empirical methods for adjusting standards:

"*They include the more technical methods and convey the impression that standards can be determined with scientific precision. In truth, however, these approaches do not determine a standard; rather, they presume that a standard already exists on an external criterion and merely translate this into a cutoff score on the test.*"

This verdict by Shepard is strongly underscored by Van der Linden in his review of decison models:

"*(...) it is emphasized that in what follows two different*

cutoff scores are involved the true and the observed cutoff scores. Decision theory can not be used to set the former; it can be used to set the latter after a solution to the former has been obtained. (...) Though this has not always been seen (e.g. Glass, 1978) the decisiontheoretic approach to ctterion referenced testing is thus no standardsetting technique but a technique to minimize the consequences of measurement and sampling error, which, preferably as a part of the normal routine, ought to follow each time a standardsetting technique is used."

Van der Linden can be shown to be wrong in his delegating the decision-theoretic approach a role of only secondary importance. What reason could Van der Linden have for his opinion? Surely he presented no mathematical evidence for it. Probably it is a particular 'Picture' of domainreferenced measurement that plays havoc with conceptual clarity. Van der Linden formulates his Picture on the same page (1980, p. 470) as follows:

"*One of the principal uses of criterion-referenced measurement is in the assignment of students to mastery states. Typically, this involves the selection of a cutoff score on the criterion-referenced scale T. Students with true scores exceeding this cutoff score are considered masters; they are deemed to have reached the learning objectives and may proceed with the next unit or task. Students below this cutoff score are the non-masters; usually, they are*

provided with extra learning time or remedial teaching."

This Picture is adequate as regards the cases where threshold utility can be used, i.e. where there is a sharp and clear demarcation between (the 'states' of) mastery and nonmastery. But Van der Linden, and of course he is not the only one to do so, stretches the Picture to include the cases where mastery is less an 'all-or-nothing' state than a gradual attribute. These are exactly the cases where linear or other contiunous utility functions on the domain score scale are appropriate.

---

figure 1 here approximately [in dit manuscript geen figuur bijgesloten, maar zie manuscript hierbeneden voor 2 figuren]

Digression on linear utility

Domain-referenced tests are used in an educational setting where the relevant outcome of instruction is understood to be the domain score or level of mastery of the student. Usually the relation between a domain score and its evaluation will be a positive one: the higher domain score is evaluated as having more utility than the lower domain score. Of course, this is an elliptical way of speaking: what is evaluated is the domain score belonging to the student who completed the course. In the following, the

fundamentally discrete character of the domain score scale will be disregarded in order to be able to use continuous utility functions on the domain score scale. In this broad class of utility functions the linear utility function is the most simple one, its general form beingu=a+bπ, where π is the domain score. Because the decision maker is free in choosing his utility scale, he may specify it so that utility zero is assigned to the domain score of zero, and utility one is assigned to the domain score of one. Figure lb pictures this utility function u=π.

The linear utility function is appropriate when (differences in) domain scores are evaluated proportionally. A domain of test-items concerning rather disconnected facts from the subjectmatter of teaching might give rise to a linear utility function. But also the domain consisting of rather difficult problems might invite the decision maker to use the linear utility function. I mention these examples only to indicate the nature of the linear utility function. In actual practice, utility functions are not determined in the rather off-hand waythese examples seem to suggest. The decision maker will have to use particular rating techniques (see Keeney and Raiffa 1976, Novick and Lindley 1979, among others) to establish three or more points of his utility curve, then the next step is to fit a linear or other continuous function on these points.

Note that a pre-established true cutoff score is not assumed here.

Utility as function of (the distance from) the true cutoff score

Threshold utility is paradigmatic whenever authors are concerned to differentiate masters from non-masters, at least until second thoughts begin to trouble the researcher's mind. Van der Linden and Mellenbergh (1977):

"*An obvious disadvantage of the (threshold loss function) is that the loss is constant. For instance, a notaccepted examinee with a latent score just above the cutting score gives the same loss as a notaccepted examinee with a latent score far above the cutting score. This constant loss can be eliminated by using a linear loss function*."

The term 'loss' as used here, as well as in Van der Linden (1980), designates negative utility, i.e. the negative of what usually is called (terminal) utility. The latent score mentioned in the quotation above refers to the domain score.

The argument of the quotation above rests on the assumption of the pre-established true cutoff score (called the cutting point). Now establishing a true cutoff score is equivalent to assigning a threshold utility function on the domain score scale. How is the assignment of a linear utility function to be squared with a pre-established true cutoff score? It simply can't be done, either you assign a threshold utility function, or you assign a linear utility function. Only the threshold utility function is consistent with the

concept of a true cutoff score. The true cutoff score has no special meaning in the case where a linear utility function is assigned, it is a score just like any other domain score. Any special meaning of the true cutoff score ought to be reflected in the shape of the utility function, but the linear function treats all domain scores equally. This particular conceptual issue, one of the main themes of my paper, can be illustrated by analyzing the mathematical form of the linear utility function as presented by Van der Linden and Mellenbergh.

The linear utility function presented by Van der Linden and Mellenbergh is a straightforward mathematical translation of the quotation above: $u=a + b(\pi-d)$, where d is the true cutoff score. (Van der Linden and Mellenbergh present this equation as the terminal utility function related to the decision to pass the examinee. However, the latter function is equal to the utility function on the domain score scale, as will be demonstrated in the next paragraph). Wilbrink (1980a) remarked that this linear function can be written as $u=a' + b\pi$, where the new constant $a' = a - bd$. Thus the true cutoff score d is seen to have no special (mathematical) meaning in this linear utility function. Nevertheless, the presentation of Van der Linden and Mellenbergh marks the origin of the current opinion that decision models assume pre-established cutoff scores.

Utility proper, terminal utility, and decision

alternatives

The cutoff score problem is to decide which observable score out of a limited number of possible scores is to be designated the cutoff score. However, the solution to this problem is to be derived from the solution of the more tractable problem of finding the best decision concerning a randomly chosen examinee. The decision alternatives in the latter problem are usually called the decision to retain or pass the student. This is a highly suggestive and misleading way of speaking, however. I will replace these terms by the more neutral 'below' and 'above' respectively. Below or above what? Not below or above the true cutoff score, as seemed to be implicit in the terms 'pass' or 'fail'. Below or above refers to the yet to be found cutoff score on the test. The decision below is the decision that this randomly chosen student, given his observed test score, has an observed score that is below the yet to be found cutoff score on the test. The decision above is the decision that the observed score of this randomly chosen student is at least equal to the yet to be found cutoff score.

Authors presenting the decision-theoretic approach to the cutoff score problem typically make use of two utility functions, corresponding to each of the decision alternatives 'below' or 'above'. These functions indicate the utility that is cashed in on or entailed by the particular decision taken. For example, when the decision is 'above', implying that the examinee concerned will not receive remedial

instruction, the utility entailed by this decision equals the utility corresponding to the domain score of this examinee. In this particular case the terminal utility function corresponding to the 'above' decision is equal to the utility function on the domain score scale. In other words: terminal utility here is equal to utility proper. The reason for distinguishing between terminal utility and utility proper is, of course, that they may be different, as indeed they are for the decision alternative 'below'.

==========

[Nederlandse vertaling van (een deel van) Passing scores]

Lineaire utiliteit

Een lineaire utiliteitsfunktie over domeinscores drukt uit dat de Waardering proportioneel is aan het percentage geweten vragen in dat domein. De algemene vorm van de lineaire utiliteitsfunktie is $a\pi + b$. Bij de keuze van de schaal voor utiliteiten kunnen we gebruik maken van het feit dat de optimale beslissing niet afhangt van een lineaire transformatie van deze schaal. Kiest men een lineaire utiliteitsfunktie, dan kan deze als zo eenvoudig mogelijk worden gekozen:

$u = \pi$.
[1]

Men wint niets aan algemeenheid door $a\pi + b$ te gebruiken, integendeel: het verdere wiskundige apparaat wordt er alleen maar complexer door.

De traditionele formule om voor raadkansen te corrigeren is een lineaire transformatie van scores, zodat funktie (1) ook is te gebruiken bij domeinscores die voor raden zijn gecorrigeerd.

Uitkomstutiliteit. Voor een doorgelaten student worden geen verdere kosten gemaakt. De uitkomstutiliteit voor deze beslissing is gelijk aan de utiliteit die is toegekend aan de domeinscore voor deze student. De funktie voor uitkomst-utiliteiten bij de beslissing 'doorlaten' is

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

waar het subscript t aangeeft dat het om uitkomst-utiliteit gaat, p staat voor de beslissing 'doorlaten', en π de domeinscore is.

Voor de beslissing 'bijspijkeren' blijft hetzelfde criterium gelden: de bereikte domeinscore (stofbeheersing), maar in dit geval na bijspijkeren. De utiliteitsfunktie over domeinscores verandert daar niet door, funktie (1) blijft van toepassing. Wat wel verandert: er zijn kosten voor het verzorgen van het

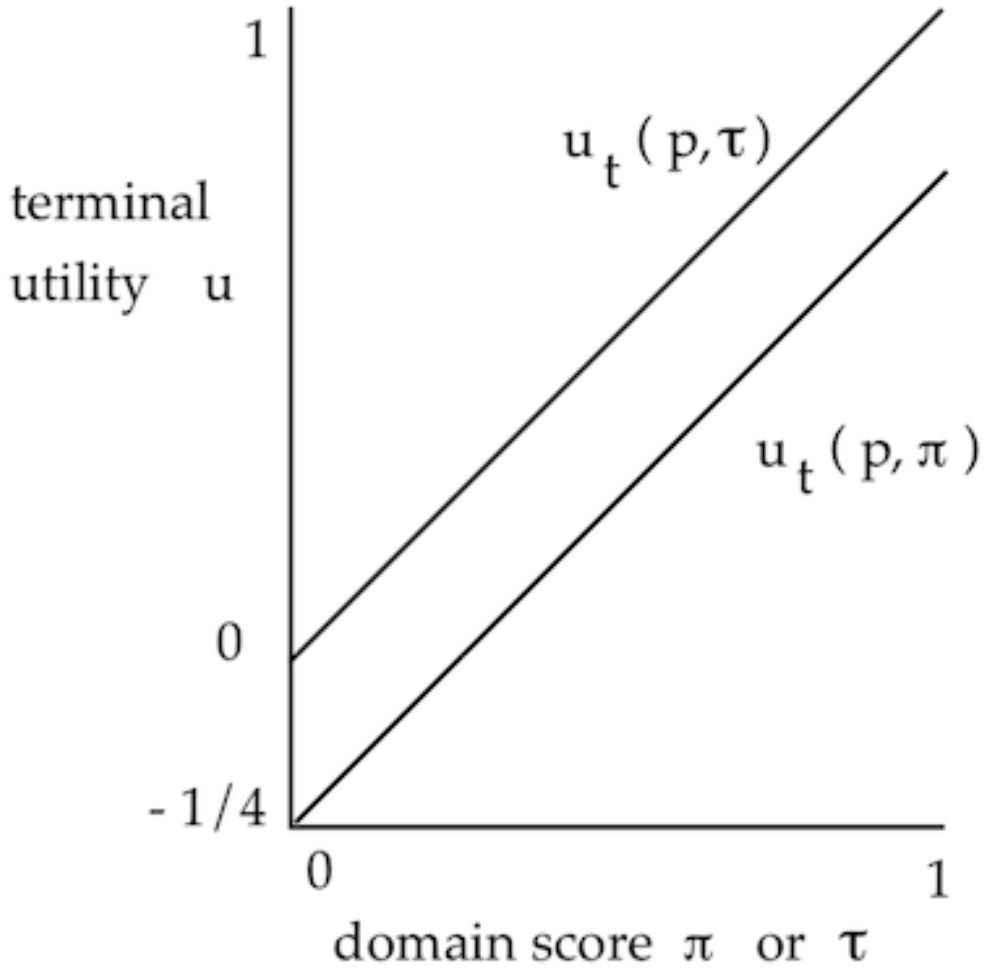
bijspijkeronderwijs, extra toetsafname, terwijl ook de student extra tijd heeft te besteden. Zolang het tegendeel niet aantoonbaar is, is het verstandig aan te nemen dat de totale kosten niet afhangen van de domeinscore van de individuele student. Omdat extreme domeinscores weinig of geen invloed hebben op de bepaling van de grensskore, is het reeds voldoende wanneer de kosten constant verondersteld mogen worden in het gebied rond de grensskore. De funktie voor uitkomst-utiliteiten bij de beslissing 'bijspijkeren' is dan:

$u_t(r,\tau) = t + c,$

[3]

[nb: + c, omdat c negatief is]

waar r staat voor de beslissing 'bijspijkeren', en τ de domeinscore na bijspijkeren is. De kosten c worden geëvalueerd op de inmiddels vastgelegde utiliteitsschaal. Om te beginnen zou men als ruwe indicatie kunnen nemen de omvang van de kosten van bijspijkeronderwijs in vergelijking tot de kosten van het reguliere onderwijs. Zou dat bijv. uitkomen op 1/4, dan krijgt c in relatie tot de al gekozen utiliteitsschaal van 0 tot 1 de waarde -1/4 (negatief, omdat het om kosten gaat). Zie figuur (1).

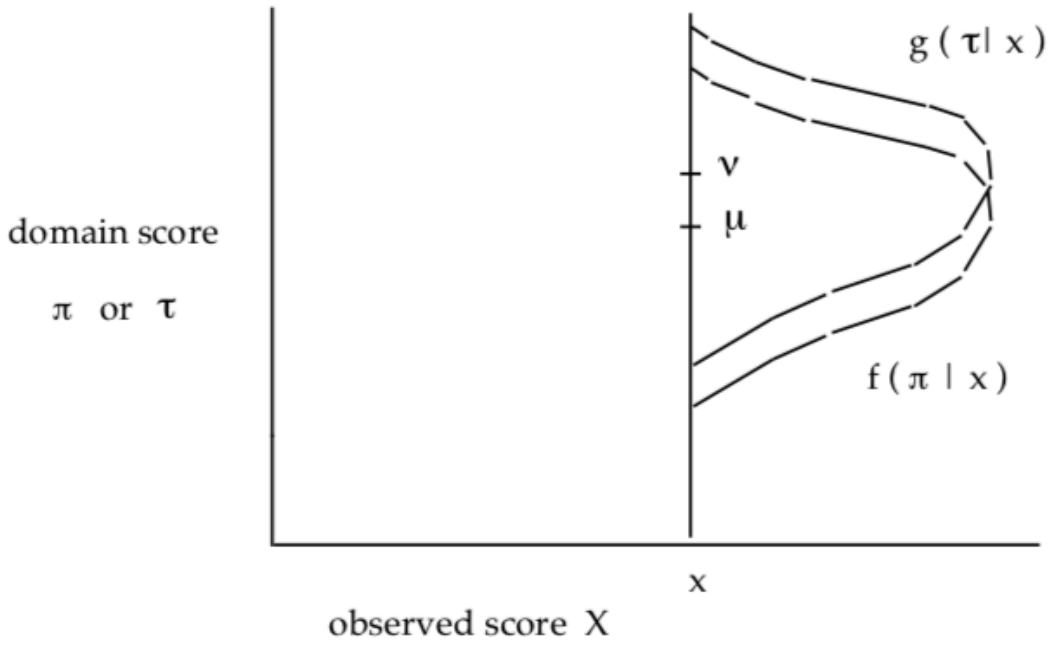


FIGUUR 1. Functies voor uitkomstutiliteiten,ut(p, π) = π voor 'doorlatenut(r, τ) = τ + c voor 'bijspijkeren'; c =.-0,25 zijn de verwachte kosten voor bijspijkeren, gegeven X=x.

[in figuur 'terinal utility'= 'uitkomstutiliteit' en 'doain score' = 'domeinscore']

Opmerkelijk is dat de functies van uitkomst-utiliteiten elkaar niet snijden. Veel auteurs veronderstellen dat deze functies elkaar wel moeten snijden, wil er een optimale grensskore gevonden kunnen worden, en

construeren de functies dan ook volgens die aanname. De fout in de laatste gedachtegang is dat het effect van bijspijkeren impliciet opgenomen is in de functie voor uitkomst-utiliteiten. In de hier gepresenteerde analyse blijft het effect dat bijspijkeren heeft op de domeinscore (de winst in stofbeheersing) strikt gescheiden van de utiliteit van de domeinscore-na-bijspijkeren. Op dezelfde wijze als de conditionele verdeling voor domeinscores is te bepalen, is ook de conditionele verdeling voor domeinscores-na-bijspijkeren te bepalen, conditioneel op **dezelfde** toetsscore X=x. Zie figuur (2).



FIGUUR 2. De conditionele waarschijnlijkheidsverdelingen over domeinscores, resp. voor bijspijkeren: f(π|x) en na bijspijkeren

$g(\tau|x)$.

Verwachte uitkomstutiliteit. Zij $f(\pi|x)$ de conditionele waarschijnlijkheidsverdeling voor de domeinscore, gegeven de waargenomen score X=x. De verwachte uitkomstutiliteit voor een beslissing gegeven X=x is te verkrijgen door voor iedere domeinscore het product te nemen van zijn waarschijnlijkheid en zijn uitkomstutiliteit, en deze te sommeren over domeinscores. Voor de beslissing 'doorlaten' is de verwachte uitkomstutiliteit:

$$E_{\pi|x}\, u_t(p,\pi) = {}_0\int^1 \pi\, f(\pi|x)\, d\pi = \mu, \qquad [4]$$

waar het subscript $\pi|x$ aangeeft dat de verwachting genomen wordt met
betrekking tot f $(\pi|x)$, en waar p de conditionele verwachte domeinscore is.
Onder de veronderstelling dat de regressie van domeinscores op waargenomen
scores lineair is, kan μ geschat worden door:

$$\mu = \rho_{xx'} \,.\, x/n + (1 - \rho_{xx'}) \,.\, \bar{x}/n \,, \qquad [5]$$

waar ρ een geschikte betrouwbaarheidscoefficiënt is, en $\bar{x}$ de gemiddelde waargenomen score.

Zij g(t|x) de conditionele waarschijnlijkheidsverdeling voor de domeinscore-na-bijspijkeren, gegeven de waargenomen score X=x. Deze domeinscorena-

bijspijkeren wordt **voorspeld** door x. De verwachte uitkomstutiliteit voor de beslissing 'bijspijkeren' is:

$$E_{\tau|x}\ u_t\ (r,\tau) = {}_0\int^1 (c + \tau)\ g(\tau|x)\ d\tau = \nu + c, \quad [6]$$

waar ν staat voor de verwachte domeinscore-na-bijspijkeren, gegeven X=x. Om een schatting voor ν te krijgen is een valideringsonderzoek nodig. Wanneer een ongeselecteerde groep studenten de behandeling 'bijspijkeren' krijgt, kan E(y|x) gebruikt worden als schatting voor ν, waar Y de waargenomen score is op de toets die na bijspijkeren wordt afgenomen. Immers, wanneer aangenomen mag worden dat E (ε|x) = 0, waar ε de meetfout is, dan is
$E_{\tau|x} = E_{(y-\varepsilon)|x} = E_{y|x}$. Onder de veronderstelling dat de regressie van Y op X lineair is, is het resultaat:

$$\nu = E_{y|x} = \bar{y}/n + (x - \bar{x})\ .\ r_{xy}\ s_y\ /s_x\ .\ n, \quad [7]$$

waar r de waargenomen correlatie tussen X en Y is, en s_x en s_y de standaardafwijkingen zijn.

Beslisregel. De beslisregel is: kies het alternatief met de hoogste verwachte uitkomstutiliteit. Vergelijk de vergelijkingen (4) en (6), en laat de student met score X=x door wanneer:

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

i.e. wanneer de kosten van bijspijkeren groter zijn dan de verwachte vooruitgang in stofbeheersing. Uit deze beslisregel volgt de optimale grensscore.

Optimale grensscore. De optimale grensskore is te vinden door voor iedere toetsscore de beslisregel toe te passen. De optimale grensscore is ook analytisch te vinden. Uit de beslisregel volgt dat de optimale grensscore die toetsscore is waarbij de verwachte uitkomstutiliteit van beide beslissingsalternatieven gelijk is, i.e. waar het de beslisser onverschillig is of studenten doorgaan, danwel worden bijgespijkerd. De optimale grensscore q is de oplossing uit:

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

afgezien van het discrete karakter van X. De berekening van verwachte uitkomstutiliteiten voor enkele geschikt gekozen waarden van X geeft het omslagpunt; de optimale grensscore q is de ruwe score die gelijk is aan dit punt, of daar juist boven ligt.

Beta-binomiaal model

Bij lineaire utiliteit is-het niet noodzakelijk de

waarschijnlijkheids-verdelingen f(π|x) en g(τ|x) nader te specificeren, omdat beslissingen slechts van de verwachte waarden en de aanname van lineaire regressie afhangen. Kies andere utiliteitsfuncties, dan zal tevens een model voor de relatie tussen domeinscores en ruwe scores gekozen moeten worden. Een bruikbaar en sterk model is het beta-binomiaal model: gegeven de domeinscore zijn toetsscores binomiaal verdeeld, terwijl de waarschijnlijkheidsverdeling over domeinscores een betaverdeling is. Onder deze aannamen is de waarschijnlijkheidsverdeling voor toetsscores de beta-binomiale verdeling. ,De parameters a en b worden geschat met door:

$$b = \frac{s^2 - \bar{x}\,(n - \bar{x})}{\bar{x} - ns^2 / (n - \bar{x})}, \quad [10]$$

$$a = b\bar{x} / (n - \bar{x}) \quad [11]$$

waar n het aantal toetsvragen is, $\bar{x}$ het gemiddelde, en s de standaardafwijking van de waargenomen scores is. De functie zelf is voor de berekeningen niet nodig, maar kan wel worden geplot om te zien of de verdeling past. De functie is term voor term te berekenen uit:

$$h(x) = \beta b(a,b,n) = (n! / [x!(n-x)!]\ )\ B^{-1}(a,b)\ B(a+x,b+n-x), \quad [12]$$

waar B(a,b) = (a-1)! (b-1)! / (a+b-1)!.

Onder de aanname van lineaire regressie van domeinscores op waargenomen scores, is de waarschijnlijkheidsverdeling voor domeinscores de betafunctie:

$$f(\pi) = \beta(a,b) = B^{-1}(a,b)\ \pi^{a-1}\ (1-\pi)^{b-1}.$$

[13]

De waarschijnlijkheidsverdeling voor waargenomen scores gegeven de domeinscore is de binomiaalverdeling:

$$h(x|\pi) = (n! / [x!(n-x)!])\ \pi^{x}\ (1-\pi)^{n-x}.$$

[14]

Het gaat echter om de verdeling $f(\pi|x)$. Gebruik makend van een bekende relatie, kan aangetoond worden dat:

$$f(\pi|x) = f(\pi)\ h(x|\pi) / h(x) =$$

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

[15]

Vergelijking (15) is gelijk aan de betaverdeling β(a+x, b+n-x).

Gemiddelde en variantie van β(a,b) zijn

a/(a+b) en ab / ((a+b)^2 (a+b+1)).
[16]

Bij berekeningen is het r-de moment rond nul van de betaverdeling nodig:

μ_r' = B^{-1}(a, b) B(a+r, b). [17]

Voor de beta-binomiaal verdeling βb(a,b,n) zijn gemiddelde en variantie:

na / (a+b) en
nab(a+b+n) / [(a+b)^2 (a+b+1)].
[18]

Bij het evalueren van verwachte uitkomstutiliteiten kan gebruik worden gemaakt van de volgende eigenschap:

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

[19]

Voorspellen van domeinscores T

Een valideringsonderzoek levert voor een ongeselecteerde groep studenten zowel de toetsscores X op, als de scores Y op een toets die na bijspijkeren wordt afgenomen. Voor beide

datasets kan het beta-binomiaal model gebruikt worden om de verdeling van domeinscores te vinden. Wat nu nog ontbreekt is een methode om domeinscores T (na bijspijkeren) te voorspellen op grond van X. Gezien de uiteenlopende aard van toepassingssituaties zal hiervoor niet een bepaald model worden ontwikkeld, maar een verdelingsvrije methode, althans wat de relatie tussen scores X en Y betreft.

Gevraagd is de voorspellende verdeling g(T|x) te bepalen. Daartoe is g(Y|x) de eerste stap. De tabel van scores X tegen Y is beschikbaar, daaruit kan voor iedere X=x berekend worden: E(Y|x) en V(Y|x), resp. het gemiddelde en de variantie van Y gegeven x. De totale variantie van Y kan als volgt uitgesplitst worden (Novick & Jackson, 1974, vergelijking 4-1.9):

V(Y) = V(E(Y|x) + E(V(Y|x).

[20]

De tweede component van (20), de variantie binnen groepen, is het deel van de variantie in Y dat niet voorspelbaar is door X.

Aangenomen is dat de varianties V(Y|x) aan elkaar gelijk zijn, i.e. gelijk E(V(Y|x) zijn (homoscedastisch zijn). Het is niet nodig aan te nemen dat de regressie van Y op X lineair is; eventueel is een vereffeningstechniek te gebruiken om grillige waarden voor E(Y|x) te voorkomen (zie bijv. Novick

& Jackson 1974, par. 10.9).

Wanneer het beta-binomiaal model past op de scores Y, past het waarschijnlijk ook op de conditionele ruwe score verdelingen, zij het dat het kleinere aantal waarnemingen mogelijk tot complicaties leidt. De verdere berekeningen zijn uit te voeren zoals in de paragraaf over het beta-binomiaal model beschreven. Op iedere g(Y|x) met gemiddelde E(Y|x) en variantie E(V(Y|x) worden de parameters van de beta-binomiaal-fit geschat; daardoor is tevens de betaverdeling g(T|x), die dezelfde parameters a en b heeft, bepaald.

attenuatie. De scores op de toets-na-bijspijkeren worden slechts verzameld om domeinscores T te kunnen schatten. Wanneer domeinscores T voorspeld worden met behulp van gegevens over toetsscores Y, wordt de voorspelling verzwakt door onbetrouwbaarheid in de toets-na-bijspijkeren. Zou met lineaire utiliteit gewerkt,worden, dan zou de correlatie tussen X en T op bekende wijze te corrigeren zijn voor attenuatie in de scores Y. Iets dergelijks is denkbaar voor de methode in deze paragraaf besproken: het gaat er dan om de tussen groepen variantie zoveel te vergroten ten koste van de binnen groepen variantie (zie vergelijking 20) dat voor onbetrouwbaarheid van de gebruikte toets-na-bijspijkeren wordt gecorrigeerd. Er is bovendien een geschikte maat voor betrouwbaarheid beschikbaar, wanneer met het beta-binomiaal model wordt

gewerkt. De betrouwbaarheid, of correlatie tussen willekeurig getrokken parallel toetsen, ofwel de Kuder Richardson formule 21, is in het betabinomiale model:

$\rho_{yy'} = n / (n+c+d)$
[21]

waar c en d de parameters zijn van de beta-binomiaal functie die op de waargenomen scores Y is gefit. Zie Lord en Novick 1968 vergelijking 23.6.14, welke vergelijking na gebruik van vergelijking (18) en na correctie van de nogal afwijkende notatie van Lord en Novick, gelijk is aan vergelijking (21).

Wanneer is te voorzien dat de optimale grensscore dicht bij $\bar{x}$ komt te liggen, dan is de attenuatie in scores Y van weinig of geen belang, en kan een correctie achterwege blijven. Bij meer extreme optimale grensscores gaat de attenuatie in scores Y echter een steeds grotere vertekening in uitkomstutiliteiten opleveren, en is correctie gewenst. Men kan ook kiezen voor een aantal toetsvragen dat in vergelijking tot dat in de 'voorspellende' toets veel groter is, althans bij de uitvoering van het valideringsonderzoek.

Omdat voorsepllingen berusten op toetsscores X, mogen deze niet voor attenuatie gecorrigeerd worden.

Drempel utiliteit

Een drempel-utiliteitsfunctie over domeinscores heeft de volgende vorm:

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

waar γ de drempel op de schaal voor domeinscores is. Er kunnen verschillende redenen zijn om voor drempel-utiliteit te kiezen. Men kan menen dat er een duidelijke grens is aan te wijzen waarboven er sprake is van 'beheersing' van de stof, en waarbeneden er niet van 'beheersing' gesproken kan worden. Of men hanteert drempelverlies als benadering voor voor andere functies die rond de 'drempel' sterk stijgen.

De uitkomstutiliteit voor de beslissing 'doorlaten' is gelijk aan verge-
lijking (28) omdat er geen andere utiliteiten of kosten zijn:

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

De uitkomstutiliteit voor de beslissing 'bijspijkeren' is:

$$-c \qquad \text{voor} \qquad \tau < \gamma$$

$$u_t(r, \tau) = \begin{cases} \\ \end{cases}$$

[30]

$$1\text{-}c \qquad \text{voor} \qquad \tau \geq \gamma .$$

waar c staat voor de kosten van bijspijkeren.

De verwachte uitkomstutiliteiten zijn

$$E_{\pi|x}\ u_t\ (p,\pi) \quad = \quad {}_{\gamma}\int^{1} \quad f\ (\pi|x) \quad d\pi \quad = \quad P_x$$

[31]

$$E_{\tau|x}\ u_t\ (r,\tau) \quad = \quad {}_{0}\int^{\gamma} \text{-c}\ g(t|x)\ d\tau + {}_{\gamma}\int^{1} (1\text{-c})\ g(t|x)\ d\tau$$

$$= \text{-c}\ (1\text{-}Q_x) + (1\text{-c})\ Q_x \quad = \quad Q_x - c,$$

[32]

waar P_x en Q_x de respectievelijke integralen aanduiden. Bij gebruik van het betabinomiale model is g(T|x) een beta-functie.

De optimale grensscore q is de waargenomen score X=x die de vergelijkingen aan elkaar gelijk maakt (afgezien van het discrete karakter van X), of waar:

$$Q_x - P_x = c.$$

[33]

Merk op dat P_x en Q_x de waarschijnlijkheid aangeven

dat de domeinscore boven de drempel ligt, gegeven X=x, respectievelijk voor en na bijspijkeren.

De hier beschreven techniek is ook te gebruiken bij andere utiliteitsfuncties die samengesteld zijn uit delen van verschillende functies. Bijvoorbeeld:

$$u = \begin{cases} \pi / \gamma & \text{voor} \quad \pi < \gamma \\ 1 & \text{voor} \quad \pi \geq \gamma \, . \end{cases}$$

[34]

PM Voor literatuurlijst zie 'Passing scores'