Multiple Evaluation versus Multiple Choice as Testing Paradigm

Feasibility, Reliability and Validity in practice

Paul Holmes

Samenstelling promotiecommissie

Voorzitter / secretaris	Prof. dr. J.M. Pieters
Promotor	Prof. dr. N.D. Verhelst
Assistent promotor	Prof. dr. A. Dirkzwager
Leden	Prof. dr. W.J. van der Linden
	Prof. dr. C.A.W. Glas
	Prof. dr. G.J. Mellenbergh
Deskundige	dr. G. Maris

ISBN: 90 - 365 17 57 - 5

Omslagontwerp: Meriel Verheul

Druk: Drukkerij Bariet bv, Ruinen

MULTIPLE EVALUATION VERSUS MULTIPLE CHOICE
AS TESTING PARADIGM –
FEASIBILITY, RELIABILITY AND VALIDITY IN PRACTICE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 7 juni 2002 te 16.45 uur.

door

Paul Holmes

geboren op 15 maart 1944

te Epsom (UK)

Dit proefschrift is goedgekeurd door:

Promotor:

Prof. dr. N.D. Verhelst

Assistent promotor:

Prof. dr. A. Dirkzwager

Acknowledgments

This thesis is based on research that was sparked by stimulating and sometimes heated debates with Arie Dirkzwager, conducted in late 1997 and early 1998 via an email discussion list. The initial plans for empirical research were made in 1999, but it wasn't until the autumn semester 2000 that an opportunity arose to put these ideas into practice.

I am particularly grateful to Arie for his continuing support and coaching throughout these years. I also wish to thank both him and Norman Verhelst for closely reviewing the preliminary versions of this manuscript. We spent many long and pleasant hours in the hospitality of Norman's home, going over every detail in an attempt to find any possible errors or unwarranted assumptions.

Apologies are due to the pilot group of students. During the first few weeks, while I was still fine-tuning the 'correction for lack of realism' and the score conversion rule, they were sometimes confronted with frighteningly low grades. As promised, however, these initial scores were subsequently recalculated. By mid-semester they already indicated cautious approval, for which I was grateful.

Over the years many others, colleagues and email contacts, have helped to clarify various details. Whether they checked the mathematical derivations, pointed out possible areas for criticism, or provided relevant references to the literature: they all helped in some way, and I wish to thank them.

Last but not least, I wish to thank my family and friends for their support. During the final stages of writing this thesis, especially, several conflicting demands were made on my time. Thank you, Hanny, for your understanding, support and patience during these months.

Summary

Multiple Choice (‘Pick One’) is widely used for testing. The scoring is rapid, objective, and fairly easy to automate. These are distinct advantages when testing large groups of students. Unfortunately, this system is flawed by the possibility of guessing. Multiple Choice with Correction-for-guessing (‘formula scoring’) was proposed as early as 1920, and has been consistently criticized ever since.

Confidence-Weighted scoring was proposed in the early 1930's. Requiring the student to indicate his confidence in his answer and weighting the item score on this basis was found to increase the overall test reliability. The results for test validity vary: some researchers report an improvement relative to Multiple Choice, whereas others report a decrease in test validity.

The next step, Probability Measurement, was proposed in the 1960's. The scoring rule guarantees that a student can maximize his expected score if and only if his response to each item honestly reflects his knowledge or skill. Nearly all research indicates that this method increases the test reliability even further. However, as with Confidence-Weighting, some researchers express doubts concerning the test validity.

This historical background is discussed in greater detail in Chapters 1 and 2.

Multiple Evaluation is a special case of Probability Measurement. The item score is based exclusively on the student’s response to the correct alternative, and this item score is calculated according to a logarithmic rule proposed by SHUFORD (1966). This rule is explained in Chapter 3, with an extension (a ‘tolerance parameter’) proposed by Dirkzwager (1997).

For practical applications in a regular educational environment some further development is required. Chapter 4 discusses the choice of ‘tolerance parameter’, conversion of the original ME test scores to standard grades, estimation of the student’s degree of ‘realism’ and ways in which the final test scores can be corrected for lack of realism. It is to be expected that this correction will reduce the possible effect of personality factors, thereby increasing the test validity. For the purposes of our research, we also consider ways of deriving equivalent estimated

Multiple Choice scores from the Multiple Evaluation responses. Last but not least, we decide how to calculate an estimate of the test reliability.

Chapter 5 presents the experimental results. In the autumn 2000 semester, first-year students taking a course in Computer Systems were presented with weekly Multiple Evaluation tests. Initially 73 students took part, but 13 dropped out of the course for various reasons. In the first week, the principles of ME testing were explained and students practised with the aid of 'dummy' (general knowledge) tests. Halfway through the course and at the end of the semester one week was dedicated to evaluation of ME versus MC by means of a questionnaire, and one other week was used for other work. In the course of the semester, this left 10 weeks for actual ME tests. During each 50-minute period the students completed two very short tests: five 4-option items by computer and ten 2-option items on paper.

Reliability estimates for the five-item 4-option tests varied from 0.58 to 0.75, as opposed to estimated MC reliabilities from 0.15 to 0.54. The ten-item True/False tests showed a similar improvement in reliability. In some cases one or more items were omitted, on the basis of item evaluation; after this correction, the ME test reliabilities varied from 0.59 to 0.90 as opposed to 0.23 to 0.65 for MC.

In line with earlier research, the students' realism was found to improve over the course of the semester. In the questionnaires the students themselves also report that they "soon learned to give realistic estimates of their personal probabilities". They also agree with the statement that "ME gives fairer and more accurate scores than MC".

The test validity was estimated on the basis of a comparison with other courses. The results of a principle component analysis indicate that the tests are valid, and that the results are not contaminated by (lack of) 'realism'.

The score conversion rule (from ME scores to regular grades) proved fairly satisfactory. There is room for improvement, however, and a new version is given in Chapter 6. Decision accuracy and measurement accuracy are also discussed in this chapter.

Based on the results reported in this thesis, we may conclude that Multiple Evaluation testing is feasible in practice, and significantly more reliable than MC.

Samenvatting

‘Meerkeuze’ is een veel toegepaste toetsvorm. De scoring is snel, objectief, en vrij gemakkelijk te automatiseren. Dit zijn duidelijke voordelen bij het toetsen van grote groepen. Helaas biedt dit systeem ook een duidelijke mogelijkheid tot gokken. Meerkeuzetoetsing met ‘gokcorrectie’ is al in 1920 voorgesteld en sindsdien met regelmatige tussenpozen verworpen, op meerdere gronden.

Ruim 10 jaar later is ‘zekerheids-scoring’ voorgesteld. De item-scores worden hierbij gewogen op basis van de mate van zekerheid die de student opgeeft ten aanzien van de juistheid van zijn antwoord. De betrouwbaarheid van de toets blijkt hierdoor toe te nemen. De gerapporteerde resultaten voor toets-validiteit zijn wisselend: sommige onderzoekers vinden een verbetering van de validiteit in vergelijking met meerkeuze, en anderen vinden juist een verslechtering.

De volgende stap, ‘probability measurement’, is in de zestiger jaren opgekomen. De toegepaste scoringsregel geeft de student een maximale verwachte item-score als zijn antwoorden een eerlijke weerspiegeling zijn van zijn relevante kennis en vaardigheid. De meeste onderzoeken tonen aan dat de toetsbetrouwbaarheid hierdoor nog verder toeneemt. De keerzijde is dat sommige onderzoekers de toetsvaliditeit in twijfel trekken, evenals bij zekerheids-scoring.

Dit eerder onderzoek wordt uitvoeriger besproken in Hoofdstukken 1en 2.

Multiple Evaluation is een specifieke variant op Probability Measurement. De item-scores hangen uitsluitend af van de opgegeven respons voor het juiste alternatief, en worden bepaald op basis van een logaritmische scoringsregel zoals voorgesteld door SHUFORD (1966). Dit is in Hoofdstuk 3 toegelicht, met een door DIRKZWAGER (1997) voorgestelde uitbreiding (een ‘tolerantie parameter’).

Praktische toepassing in een reguliere onderwijsomgeving vereist nog enkele aanpassingen. Hoofdstuk 4 bespreekt de keuze van de tolerantie parameter, de omzetting van ruwe ME scores naar gebruikelijke schoolcijfers, schatting van het ‘realisme’ van de student en methoden om de totaalscore te corrigeren bij gebrek aan realisme. De aanname is dat het effect van persoonlijkheidsfactoren hierdoor afneemt, zodat de toets-validiteit verbetert. In verband met ons onderzoek worden methoden besproken om geschatte ‘meerkeuze-scores’ af te leiden uit de ME antwoorden, en methoden om de toets-betrouwbaarheid te schatten.

Hoofdstuk 5 geeft onze praktijk-resultaten. In het najaarsemester 2000 zijn de eerstejaars studenten wekelijks getoetst bij een module ‘Computer Systemen’. Van de oorspronkelijke 73 studenten zijn 13 tussentijds gestopt met hun studie. In de eerste week is het principe van ME toetsing uitgelegd, gevolgd door een praktische kennismaking via oefentoetsen. Halverwege het semester en aan het einde is steeds een week uitgetrokken voor een enquête, waarbij de studenten ME toetsing vergeleken met standaard meerkeuze (MC), en één week werd besteed aan ander werk. Gedurende het semester bleven dus 10 weken beschikbaar voor toetsing. In één lesuur (50 minuten) werden steeds twee zeer korte ME-toetsen afgenomen: vijf 4-keuze items via computer-afname, en tien 2-keuze items via papier&potlood.
De geschatte toetsbetrouwbaarheid voor de 5-item 4-keuze toetsen lag tussen 0.58 en 0.75, tegen slechts 0.15 tot 0.54 voor geschatte MC resultaten. Ook de betrouwbaarheid van 10-item Juist/Onjuist toetsen nam sterk toe. Nadat bij sommige toetsen enkele items vervielen, op basis van itemanalyse, bleek bij ME de betrouwbaarheid te liggen tussen 0.59 en 0.90, tegen 0.23 ... 0.65 voor MC.
Evenals bij eerdere onderzoeken bleek het realisme van de studenten in de loop van het semester te verbeteren. In de twee enquêtes geven de studenten aan dat zij “snel leren realistische schattingen te geven”. Ook zijn zij het eens met de stelling dat “de toetsresultaten bij ME eerlijker en juister zijn dan bij MC”.
De toets-validiteit is geschat op basis van een vergelijking met andere vakken. De resultaten van factor-analyse wettigen de veronderstelling dat de toets-scores valide zijn, en niet beinvloed door (gebrek aan) ‘realisme’.
De toegepaste scoreconversie, van ME scores naar een reguliere onderwijsschaal, blijkt al redelijk bevredigend. Toch is verbetering mogelijk, en een nieuwe opzet is in Hoofdstuk 6 gegeven. In dit hoofdstuk besteden wij ook aandacht aan de meet- en decisie-nauwkeurigheid.
Op basis van de gevonden resultaten kunnen wij concluderen dat Multiple Evaluation toetsing praktisch toepasbaar is in een reguliere onderwijsomgeving, en significant betrouwbaarder is dan traditionele meerkeuzetoetsen.

Contents

1 - Introduction

1.1 From Multiple Choice to Probability Measurement

The most widely used method for structured test questions is Multiple Choice (MC): for each test item, the candidate must pick one of several alternatives as ‘correct’. There are many variations on this basic idea, varying from statements (that may be True or False) to items where several alternative answers are presented of which only one is correct, or even items where more than one of the presented alternatives may be correct. The item score is then based on the alternative selected by the candidate: full marks if it was the correct one, or zero points (possibly even penalty points) if it was wrong. Note that in the latter case the points for the item are determined by the *incorrect* alternative selected. The candidate’s thoughts concerning the correct alternative are unknown and therefore cannot be taken into account.

However, we should realise that in all cases the candidates actually select the alternative which they consider *most likely*, on the basis of their current knowledge and interpretation of the item. They may consider a statement more or less likely to be true. When several alternatives are presented, they may consider one of these more or less likely to be the (only) correct answer, or they may consider several alternatives (almost) equally probable. Based on this personal evaluation, they must then pick one alternative as their answer.

When a candidate picks the correct answer this choice may be based on almost perfect certainty (100% knowledge), considerable uncertainty (with several alternatives appearing ‘likely’), or even a lucky guess (with no knowledge of the topic involved). On the other hand, an incorrect choice may be the result of a bad guess, uncertainty, or even serious fallacy. Traditional MC tests provide no means for discriminating between these possibilities. Furthermore, the possibility of guesswork introduces an element of chance, so that the test results will always contain some ‘random noise’ - hence the requirement for a sufficiently large number of items to allow valid statistical analysis.

We must admit, therefore, that traditional MC is a coarse measuring tool. The recorded response (alternatives marked) contains no information on how the candidate selected each particular alternative: complete certainty, vacillation between several alternatives, blind guess, or even more or less serious fallacy. We are attempting to view and measure the candidates' knowledge and capabilities through an extremely coarse filter which lumps all personal certainties, doubts, fallacies and guesswork into just two categories: 'correct' or 'incorrect'.

Several methods to overcome this problem have been proposed, as will be discussed in somewhat greater detail in Chapter 2. With only a few exceptions, these methods can be grouped into three main categories:

1. **Correction For Guessing.** In its simplest form, 1 point is awarded for each correct response and 1/(k-1) penalty points are given for each incorrect answer (where k is the number of alternatives). For True/False statements (k=2) this system is also referred to as "number right minus number wrong".
   Candidates may be encouraged (or even instructed) to "Omit, do not guess", for example by means of a True/?/False system where a '?' response gives zero points. It has been shown (as will be discussed in Section 2.1) that this system is far from ideal. If the scoring rule is symmetrical (with maximum 'correct' scores equal to maximum penalty points) students will maximize their expected test score by always making extreme 'True' or 'False' choices and ignoring all intermediate options.

2. **Confidence Weighting.** For True/False statements, a graded response may be used ("True and certain", "True and uncertain", "?", False and uncertain", False and certain", for example), where a "certain" response gives more points if correct but carries a greater penalty if incorrect. For traditional pick-one Multiple Choice items, where several alternative answers are proposed of which one is correct, an additional scale for 'certainty' may be added. Students must first choose one of the alternatives as the 'correct' answer, and then indicate on the secondary scale how sure they feel about this response. It should be noted, as several researchers have pointed out, that this secondary scale should

preferably be numerical instead of the more common ill-defined verbal scales such as ‘quite certain’, ‘fairly certain’, ‘unsure’. On such ‘fuzzy’ scales the interpretation may vary from one student to another, unless quite explicit information is given concerning the effect of the scoring rules.
This system is sometimes referred to as “weighted number right” or, when correction for guessing is included, “weighted number right minus weighted number wrong”. Although it may have some value when the correct alternative is selected, it is of doubtful worth when a distractor is chosen ‘with more or less certainty’. (See Section 2.2).

3. A third approach, **Probability Measurement**, was originally proposed in the 1960's. As with Multiple Choice, each test item is presented with a set {1, 2, ..., k} of alternative answers. Instead of picking one, however, the student must respond by assigning a personal probability r_j to each of the alternatives, where $0 \leq r_j \leq 1$ and $\Sigma r_j = 1$ for $j = 1, 2, ..., k$. The item score is based on the probability r_c assigned to the correct answer.

Although Probability Measurement bears a superficial resemblance to Confidence Weighting, there is a fundamental difference between the two concepts.
With Confidence Weighting the student must pick one of the alternative answers (as for MC) and then assign a probability to the correctness of this particular choice. If in fact an *incorrect* answer was chosen, no information is available concerning the student’s evaluation of the *correct* answer (was it also considered possible? unlikely? or even impossible?).
Probability Measurement elicits probability assignments for *all* alternative answers. This allows us to base the item score on the student’s evaluation of the correct answer, even ignoring the responses to all incorrect alternatives if we so wish.

Only for the case where k = 2 (e.g. True/False statements) can Confidence Weighting provide similar information, if we assume that the response and probability assigned to one option will usually imply the response and probability that would be assigned to the other.

1.2 From Probability Measurement to Multiple Evaluation

Probability Measurement was analysed in mathematical detail by SHUFORD, ALBERT AND MASSENGILL in 1966 [70]. In the opening paragraphs, they sum up earlier contributions as quoted in Figure 1. (Note that in this and all following quotations, the literature references have been modified to correspond to our reference numbering.).

To briefly summarise this summary: conventional Multiple Choice extracts only a very small fraction of the information potentially available. Degree-of-belief probabilities concerning the correctness of the various possible answers provide far more information, and procedures to measure these are available.

Limiting ourselves to tests in which possible answers to each question are given (as multiple alternatives for each item, as in Multiple Choice) and furthermore restricting ourselves to Symmetric Reproducing Scoring Systems (as defined by Shuford; see Section 3.1), we can consider a promising special case: Multiple Evaluation (ME).

In Multiple Evaluation tests each question is presented with a set of alternative answers, and the student must assign a personal probability to each of these alternatives. The item scores are based exclusively on the personal probability assigned to the *correct* alternative; probabilities assigned to incorrect alternatives have no effect on the item score. Assigning 100% probability to the correct alternative gives full marks; 75% on the correct alternative (and therefore 25% on all others combined) gives less than full marks. Assigning 100/**k** % (where **k** is the number of alternatives; e.g. 25% for **k**=4) is taken as 'don't know' and awarded zero points. An even lower probability assigned to the correct alternative indicates that the candidate considers this (correct) alternative improbable; this is taken as an indication of an error or misunderstanding of the subject matter, and results in points being deducted.

The item scores are calculated according to a modified logarithmic 'Proper Scoring Rule', as will be discussed later. This rule is designed to give the highest item score when the candidate is 'realistic', i.e. when he or she accurately estimates the personal probability for each alternative.

> “Upon reflection it is quite apparent that all techniques in current use for assessing the present state of a student’s knowledge fail to extract all of the potentially available information. In the case of objective testing and programmed instruction techniques, it has been argued [69] that the choice and constructed response methods upon which they are based extract only a very small fraction of the information potentially available from each query. It has been further argued [58, 69] that all of this information is contained in the student’s degree-of-belief probabilities [14, 62] concerning the correctness of the various possible answers.
>
> In order to be able to measure these degree-of-belief probabilities in an educational environment, it is necessary to have a scoring system designed in a very special manner that guarantees that any student, at whatever level of knowledge or skill, can maximize his expected score *if and only if* he honestly reflects his degree-of-belief probabilities [69]. Testing procedures which utilize such scoring systems and which make their properties known to the student will be termed *admissible probability measurement procedures*.
>
> None of the commonly used measurement procedures, e.g. direct estimation, category judgements, direct ratio scaling, and indifference procedures are admissible in the sense defined in [69]. It is important to note, therefore, that admissible probability measurement procedures do, in fact, exist. Masanao Toda, in the early 1950's, experimented with two of these procedures which he later termed [83] the *quadratic loss* and *logarithmic loss one-person games*. Van Naerssen [86] independently discovered these two scoring systems. The quadratic scoring system was independently discovered and generalized to the case of more than two alternatives by de Finetti [15]. Roby [66] discovered a third scoring system which is called the *spherical gain game* by Toda [83]. Toda reviewed these three games along with other games appropriate to the measurement of degree-of-belief probabilities distributed over a continuum and rationalized their essential character in terms of the *matching property* which requires that the optimal probability assignment be proportional to the student’s degree-of-belief probabilities.
>
> In this paper, we integrate and generalize these known results.”

Figure 1. Summary of earlier research, quoted from SHUFORD [70].

In broad terms this implies that of all alternatives assigned a probability of 75%, for example, three out of four should be correct and one out of four incorrect: the proportion correct p should be equal to the personal probability assignment r for all values of r. If, for $p > 1/\mathbf{k}$, a larger proportion is correct ($p > r$, on average) the candidate is underestimating his capabilities; conversely if a smaller proportion is correct ($p < r$) he is overestimating. A method for calculating the 'degree of realism' will be discussed later.
Given that candidates attempt to maximize their test score, a proper scoring rule encourages them to be realistic and discourages more-or-less random guesswork. This tends to eliminate the element of chance which is inherent in MC testing. Students are also encouraged to consider the item and all alternatives more carefully, and to attempt to realistically estimate their knowledge (and/or accuracy) for this topic. These three effects are illustrated in the examples given in Figure 2.

As a final educational bonus, meaningful feedback can be given: 'You are overconfident, you don't know as much as you think', or, alternatively, 'You underestimate your capabilities. You know more than you think.' This feedback is essential during the initial (learning) stage of Multiple Evaluation testing, to steer the students towards more realistic probability assignments. It also seems to have a more profound effect in some cases (although this was not studied as part of the current research), witness comments such as "If there is one thing I learned, it is to think twice in future" and, on the other hand, "Tests scare me and make me feel uncertain. I had to really push myself to go for the full 100% on just one of the alternatives."

All in all, it would seem that Multiple Evaluation is a far better testing method than Multiple Choice. However, if the basic theory has been known since the 1960's, one must wonder why it is still not common practice some forty years later?
As will be clarified in the following chapters, there are two main reasons for this lack of success: doubts regarding the practical feasibility of Multiple Evaluation testing, and doubts regarding the validity.

Q1. +5.3125 = $5\ {}^{5}/_{16}$ (decimal) is represented in IEEE floating point format as 0.10000011.01010100000000000000000. Therefore, 0.10000**100**.01010100000000000000000 in this format corresponds to:

a. $+10\ {}^{5}/_{8}$ b. $+\ 8\ {}^{5}/_{16}$ c. $+\ 2\ {}^{5}/_{8}$ d. $+\ 1\ {}^{5}/_{16}$

For a candidate who has studied the subject matter, this is a fairly easy question: the original value is multiplied by 2, so the correct answer is **a**. Without this knowledge, however, Multiple Choice tends to degenerate into Multiple Chance: pick one of the alternatives, at random in the worst case, with a 25% chance of being correct.

In an actual ME test, the correct alternative was chosen with 100% certainty by 10 out of 65 students, and with a high degree of probability (r > 50%) by a further 10. Of the remaining 45 students, only 5 clearly rejected the correct alternative (r < 12%); the rest either left the question unanswered (with r = 25% for all alternatives) or hazarded a cautious guess (12% < r < 50%).

Q2. The characteristic distinction between digital and analogue systems is:

a. digital systems provide better quality than analogue systems.

b. digital systems are less expensive than analogue systems.

c. analogue systems work with continuous signals, digital systems don't.

d. analogue systems use discrete components, digital systems use chips.

Although each of the alternatives may appear to contain an element of truth, more careful reading will show that only **c** can be considered a 'characteristic distinction'. As one of the students said afterwards: "At first I was going to pick **a**, but then I read through the whole question again, more carefully, and realized it should be **c**."

Q3. The hexadecimal number 83CF$_H$ corresponds to the binary number:

a. 10001111001111_B

b. 1000001111001111_B

c. 1000001110011111_B

d. 1000111100111100_B

For half the students, a straightforward conversion led to the correct answer: **b**. More to the point, one-third responded with r ≈ 25%, i.e. 'Don't know'. When queried about this afterwards they gave a simple reason: "I always get those calculations wrong."

Figure 2. ME tests discourage guesswork (Q1), and encourage careful consideration of the question (Q2) and realistic estimates of the students' knowledge and/or accuracy (Q3).

Many early researchers cast doubts on the practical feasibility of Multiple Evaluation testing. ME demands a fairly complicated scoring rule which leads to several practical problems both for students when taking the test and for examiners when scoring the test afterwards. It is only fair that candidates taking a test (*any* test!) are aware of the consequences of the scoring rule that is to be applied. This knowledge may be more or less 'fuzzy', as for essay-type items, or simple and clear-cut as for MC - e.g. 'full marks if correct, otherwise zero'. The far more complicated scoring rule for ME requires extensive calculation or at the very least a look-up table. At the same time, knowledge of the possible ME scores is an important aid when students must assign personal probabilities to several alternatives. For this reason, valid ME testing requires a computer-based testing environment - certainly if each item is presented with three or more alternatives. As early as 1965 SHUFORD [69] reported experiments with computer-based testing, followed in 1968 by BAKER [6], in 1974 by SIBLEY [76] and in 1975 by DIRKZWAGER [19], but it wasn't until the late 1980's that computers became available (and affordable) in sufficient quantity in practical educational environments. Even then, the introduction of ME was hampered by the lack of suitable software until TestBet© became available in 1998.
A further problem concerning the feasibility of ME testing is that the proper scoring rule required (in its basic form) leads to severe penalties for misunderstanding (theoretically, a minus-infinite item score in the worst case). SHUFORD, ALBERT AND MASSENGILL presented one possible solution as early as 1966 [70], and DIRKZWAGER proposed a more elegant solution in 1997 [22]. Even so, the possibility of extreme negative scores made it all too easy to dismiss this proper scoring rule: "This is a ridiculous system. You can't possibly give a student a score of minus infinity!".

The second main problem is that many researchers have expressed doubts concerning the validity of ME testing. For this system to work as intended, the scoring rule must reward 'realism' and penalise lack of realism. The final scores must therefore not only reflect the student's knowledge or aptitude, they must also to some extent reflect a 'personality variable'. However, it has been shown by

several researchers (notably Sibley [76] and Dirkzwager [21]) that this effect tends to disappear with practice. Most students quickly learn to give realistic responses.
To summarize: computer-based testing was too expensive or even impossible, but this problem has now been solved. The reliability of ME was never in doubt, and personality traits should not influence the validity after some initial training.
Given these facts, it is time for a re-appraisal of Multiple Evaluation testing. The purpose of our research was to evaluate ME and determine whether it could present a better testing tool than MC. In particular, we wished to evaluate the feasibility of using ME for certifying tests in a practical educational environment.

1.3 Scope of the current research

Testing can be roughly divided into three distinct steps: (1) item and test *construction*; (2) test *administration* (application and scoring); and (3) *evaluation* (test and item analysis). This is an iterative process: based on the evaluation, item and test construction can be improved for later examinations.

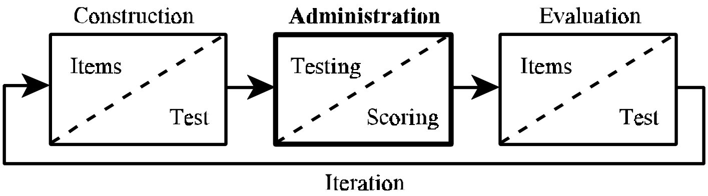


Figure 3. Three main steps in testing procedure.

In the current research we are primarily interested in the central step: Multiple Evaluation testing methods and scoring, as applied and evaluated in a practical educational environment. Item and test construction will not be discussed, although some consequences of ME testing will be noted in passing.

Although not our prime interest, item and test evaluation must be discussed in some detail if we are to evaluate the test results obtained. We may note, in passing, that ME testing provides a wealth of additional information concerning the individual items, which may prove invaluable for more detailed item analysis. However, that does not lie within the scope of our research.

1.4 Primitive notions

When students respond to a Multiple Evaluation test, several factors enter into the equation. For the purpose of this text we will distinguish the following 'primitive notions':

Personal beliefs: what the student believes to be true about the subject matter. Personal beliefs can vary from accurate knowledge, through approximation or inaccuracy, to serious misconceptions. As an example, a student may personally believe that Copenhagen is 'the capital city of Denmark', 'a city somewhere in Europe', or 'the capital city of The Netherlands'.

Degree of belief: how much confidence the student has in his 'personal belief' concerning the subject matter. In the above example, a student who correctly believes that Copenhagen is the capital city of Denmark may feel sure of this answer (because he is good in geography), or quite unsure (because he often gets this sort of question wrong). On the other hand, the student who places Copenhagen in The Netherlands may be highly unsure (fortunately so!) or quite sure - which would be a serious misconception.

True (personal) probability: based on the student's knowledge of the subject, there is a (personal) probability, p, that any particular answer to an item is correct. This should not be confused with the (factual) probability that an alternative is actually correct: obviously, just one alternative should be 100% correct for most multiple-choice items. When the student selects an alternative, however, there is a 'true, personal probability' that this *choice* will be accurate.

To quote De Finetti [18]:

> "... we shall ask ourselves the following three questions:
> - 'probability of what?', and to this we shall answer 'of an event';
>   and then
> - 'in what circumstances?', and here it is natural to answer 'taking into account all the circumstances known to be relevant at the time';

and lastly

– ‘evaluated by whom?’ to which the only possible answer is ‘by the subject considering them’.”

In our case the ‘circumstances’ are the subject’s personal beliefs and degree of belief concerning the test item under consideration, with the alternative answers proposed. For a well-constructed item, the student’s personal probability that a given alternative is correct should be a good measure of his knowledge of the subject matter.

Unfortunately, we cannot measure this personal probability directly. Instead, we must estimate it on the basis of the

Observed response: based on his or her ‘personal belief’, ‘degree of belief’ and ‘realism’ (see below), a student must rate each alternative answer to each item by means of a percentage. This response, *r*, is what the examiner observes.

For Probability Measurement in the general sense, the item score may be based on the observed responses for *all* alternatives for each item. Multiple Evaluation item scores are based exclusively on the response for the *correct* alternative.

In the above example: the student who maintains that Copenhagen is the capital city of Denmark may only rate this answer at 75% likely (observed response, *r*) although, given his knowledge of geography, his true probability *p* is closer to 100%. Conversely, a student may respond to this and similar items with 100% certainty (*r* = 100% for all these items), while only actually getting four out of five right. In that case, his true personal probability would be closer to *p* = 80%. The first student is underestimating his knowledge; the second student is overestimating. Which leads us to the next primitive notion:

Realism: simply stated, this is the degree of correspondence between the observed response *r* and the true personal probability *p*. If a student consistently underrates his knowledge and proficiency, the observed response will be less than the true personal probability: he is underestimating. Conversely, overestimators respond with more extreme percentages (closer to 0% or 100%) than is consistent with their true knowledge. Over a large test, the average ‘number right’ of all alternatives that a realistic student has rated as 75% should be 75%. If more than 75% are actually correct, the student is underestimating his capability; if less than 75 % are correct he is overestimating.

It should be noted that lack of ‘realism’ implies that the observed response will not be identical to the true personal probability. In most cases we may wish that the test results should reflect the students’ true probabilities, and for ‘realistic’ students the observed response will be an accurate measure. For less realistic students (both underestimators and overestimators) ‘realism’ can be estimated for ME tests, and some correction can be applied - as will be described.

More importantly, the scoring system is specifically designed to reward realism (as will be described in Chapter 3). Given sufficient feedback, therefore, all students in our study soon tended towards more realistic responses. This result is consistent with earlier research (see Chapter 2). Note that this is a two-pronged approach: students are steered towards realism, but while they are still insufficiently realistic the final scores are corrected to some extent.

1.5 Outline

In the following chapters we will discuss the historical and theoretical background of Multiple Evaluation, propose some further modifications to the scoring system, present the calculations used to compare ME-results with (derived) MC-scoring, and discuss the actual results obtained.

The historical background is dealt with in Chapter 2. As we shall see, Multiple Evaluation evolved as a solution to the twin problems in structured (multiple alternative) test items: insufficient information concerning the actual knowledge of a student, and the inaccuracy caused by ‘guessing’. Other partial solutions can be grouped into three main categories: Correction for Guessing, Confidence Weighting and Probability Measurement (of which ME is a special case). Early research tends to show that Probability Measurement offers the highest test reliability, but at the expense of lower test ‘efficiency’ and with some doubts cast on test validity. The problem of efficiency is easily solved, nowadays, by means of computer-based testing. Test validity is re-considered in Chapter 5.

In Chapter 3 we will discuss the theoretical background of Multiple Evaluation, relying heavily on earlier work by SHUFORD and DIRKZWAGER. It will be shown that Multiple Evaluation, using a logarithmic scoring rule, appears to be the best scoring system for structured test items.

Some practical problems must be solved, however, before this scoring system can be implemented in a regular educational environment. In Chapter 4 we will turn our attention to the ‘tolerance parameter’ initially proposed by Dirkzwager (designed to eliminate the problem of ‘minus infinity’ scores associated with the logarithmic rule). We will propose a conversion rule from ME scores to regular test scores, and methods for correcting the final test scores for ‘lack of realism’. Methods of deriving comparative MC scores, for the purposes of this research, are also discussed. Finally, the calculation method for ‘test reliability’ is described.

Chapter 5 presents the experimental results obtained. For a sufficiently reliable test, a surprisingly small number of items proved sufficient: only five to seven four-alternative items or some ten two-alternative (True/False) items. These results are discussed in some detail, and comparisons are made between the ME and (derived) MC test results. Our students were also asked to evaluate this system, as compared to ‘traditional Multiple Choice’, and their responses are favourable. Test validity is estimated on the basis of the same students’ results for some other parallel subjects, with intriguing results. Finally, some attention is paid to the ‘feedback’ given to the students, particularly concerning their ‘realism’ - or lack of it.

In general, the results of our initial experiments with Multiple Evaluation are quite promising There is still room for improvement, however. In Chapter 6 we suggest a more sophisticated score conversion rule (from ME scores to a regular test scoring scale), and its effect on decision accuracy and error of measurement.

Finally, the Appendices include information that we considered of interest in the context of our research - although not essential in the main body of the text.

1.6 Glossary and Notation

In order to maintain consistency throughout the text, the following notations for variables and parameters are defined. In quotations, all notations have been modified to correspond to those listed below and all literature reference numbers correspond to our numbering.

Variables:

i	item
j	alternative answer for item
r	(observed) response
	r_c personal probability assignment for correct alternative
	$r_c(i)$ personal probability assignment for correct alternative, item i
	$r(j)$, r_j personal probability assignment for alternative j
	$r(i,j)$ personal probability assignment for alternative j, item i
p	true personal probability
	$p(j)$ or $p(i,j)$ true personal probability for alternative j (item i)
$s(i)$	score for item i
	$s(j)$ or $s(i,j)$ item score if alternative j is correct (item i)
S	actual (observed) overall test score
$E[S]$	expected overall test score

Parameters:

k	number of alternatives per item
m	number of items in a test
n	number of students in a test group
t	tolerance parameter
T	tolerance ratio

Abbreviations:

MC	Multiple Choice (answered by selecting one of the alternatives per item)
ME	Multiple Evaluation (answered by probability assignments for each alternative for each item)
(S)RSS	(Symmetric) Reproducing Scoring System

2 - Related research

Over the last century, many papers have been published concerning the relative merits of various scoring systems for structured test questions. As stated earlier, most of these systems can be grouped into three main categories: Correction for Guessing, Confidence Weighting and Probability Measurement - with Multiple Evaluation as a special case of the latter.

In this chapter we shall review the main results and draw some tentative conclusions.

2.1 Correction for Guessing

Even if we have absolutely no knowledge of a subject, we can answer a multiple choice question by marking one of the alternatives at random - with a 1/**k** chance of being right. Over a large number of items, therefore, random guesswork will give a positive contribution to our total test score. Correction For Guessing attempts to compensate for this effect.

In its simplest form, 1 point is awarded for each correct response and 1/(**k**-1) penalty points are given for each incorrect answer (where **k** is the number of alternatives). For True/False statements (**k**=2) this formula reduces to "number right minus number wrong". What this system does, in effect, is to assume that all incorrect responses are the result of random guesswork and that therefore a corresponding proportion of all correct responses must also be the result of random guessing. For **k**=4, for example, three incorrect answers will cancel out one correct response. Candidates are usually offered a (more or less explicit) escape route: "If you do not know the answer, omit! Do not guess".

This system has been consistently criticized since 1923 (TROW, [84]). Notably, both SLAKTER (1968, [78]) and HARRIS (1969, [42]) point out that it penalizes risk-avoiders: non-guessers who choose to 'omit' rather than to answer a question when they are not sufficiently sure, for fear of the associated penalty points. Even guessing on the basis of partial knowledge is shown to lead to a higher expected score (overall) than omitting to answer questions when one is unsure. FRARY (1977, [35]) expressed this as "Correction for Guessing devaluates partial information", and LECLERCQ (1983, [55]) again proved that it is in the student's best interest to guess rather than omit. BACHMAN & PALMER (1996, [5], pp.204-205) make the distinction between 'random' and 'informed' guessing.

Going one step further, HUGHES (1989, [48]) maintains that the amount of guesswork is 'unknowable' for Multiple Choice; he concludes that both the basic system (without correction) and correction for guessing should be avoided wherever possible. BRUNO (1993, [11]) warns that the final scores for uncorrected Multiple Choice are biassed upwards (due to 'lucky hits'), but that the scores for Corrected for Guessing MC are biassed downwards (since they assume random, uninformed guessing).

In an earlier study, EBEL (1968a, [30]) specifically elicited information on the amount of blind guessing in MC tests. He found a weighted average of only 5.8%, with the caveat that the amount of guessing depends on the difficulty of the items: the most difficult items could score up to 40% blind guesses. We must conclude that a certain amount of guessing indeed occurs, but that it will usually be far less than the simply-calculated $1/\mathbf{k}$ (e.g. 25% for 4-alternative items) and will depend on the (perceived) item difficulty.

A sweeping 'Correction for Guessing' based on $1/\mathbf{k}$ will therefore be unfair to the majority of students, and offering the option to 'omit' makes things even worse.

As a final point, it should be noted that asymmetric correction-for-guessing (with more severe penalties than $1/(\mathbf{k}-1)$ for incorrect answers, in an attempt to force students to 'omit, rather than guess') cannot solve the basic problem. As a simple case, we can examine the effect when $\mathbf{k} = 2$ (e.g. True/False items). Defining a 'penalty factor', $\mathbf{f}$, as the ratio between the penalty for incorrect answers and the points gained for correct answers (so that $\mathbf{f} = 2$, for example, corresponds to -2 and $+1$ points respectively), we find that the Expected Score E(S) for each item depends on the student's personal probability as follows: $E(S) = \mathbf{p} - (1-\mathbf{p})\times\mathbf{f} = \mathbf{p}\times(1+\mathbf{f}) - \mathbf{f}$. Note that the Expected Score is zero if an item is left unanswered; *any* improvement on this is to the student's advantage. For $\mathbf{f} = 0$ (no penalty) the Expected Score equals $\mathbf{p}$; for $\mathbf{f} = 1$ (symmetric penalty) $E(S) = 2\mathbf{p} - 1$, which is greater than zero for all **p** > 0.5 – so that a student should 'guess' if he has any idea at all concerning the correct answer. For larger values of **f** (more severe penalties) E(S) is greater than zero for **p** > **f** / (1+**f**). Although students will no longer maximize their expected test score by always making extreme 'certain' choices, the best strategy is still to make a clear choice when they feel sufficiently certain. In this case, 'sufficiently' certain means that they estimate their chances at more than *x*% (where *x* depends on the scoring rule). If this is what we wish to achieve, Confidence Weighting would appear a more appropriate option.

2.2 Confidence Weighting

For True/False statements, a graded response may be used ("True, certainly", "True, probably", "?", False, probably", False, certainly", for example), where a "certain" response gives more points if correct but carries a greater penalty if incorrect. This system is sometimes referred to as "weighted number right" or, when correction for guessing is included, "weighted number right minus weighted number wrong". For True/False statements (**k** = 2) this scoring rule is actually a first approximation of Probability Measurement (which will be dealt with in the next section), since a choice with a specific degree of certainty *for* 'True' automatically implies a choice with 'one minus that degree of certainty' *against* 'False'.

For traditional pick-one Multiple Choice items, where several alternative answers are proposed of which one is correct, an additional scale for 'certainty' may be added. In that case, students must first choose one of the alternatives as the 'correct' answer and then indicate on the secondary scale how (un-)sure they feel of this response. Although this may have some value when the correct alternative is selected, it is of doubtful worth when a distractor is chosen 'with more or less certainty'. In the latter case, the student's response contains little or no information concerning his thoughts concerning the correct response. As an example, assume that 'A' is the correct response for a **k** = 4 item, but the student selects 'B' with 50% certainty. His personal probability concerning the correct answer can be anything between 50% and 0% in this case – varying from ‘equally probable’ (considerable knowledge, but an unlucky choice?) to ‘impossible’ (which might indicate a serious fallacy). In our view this type of scoring system could only have value if *all* distractors are constructed with the greatest care, so that a choice for a distractor also contains meaningful information. This may be expected to prove extremely difficult, if not impossible, in practice.

HEVNER (1932, [45]) was one of the first to evaluate this method for true/false tests, in music appreciation and art judgement. She used a three-point confidence scale, whereby students could claim 3, 2 or 1 points depending on the amount of confidence they reported. She compared four scores: traditional MC and MC with correction for guessing (i.e. ‘number right’ and ‘number right minus number wrong’), and Confidence Weighting without and with a penalty scale of –2, –1 or 0 points depending on the confidence reported. Using split-half comparisons, she found the highest reliability for ‘weighted number right’ scores. These scores also had the highest validity, using scales for music talent and music training as comparison criteria. However, she notes that the scoring rule for weighted number right must be kept secret “so that dishonest students cannot artificially raise their scores”.
A similar increase in reliability was found by SODERQUIST (1936, [79]), when comparing ‘right minus wrong’ to ‘weighted right minus weighted wrong’. In this case a 4-point confidence scale was used *and* the penalties were double the corresponding credits: 4/3/2/1 points credit as opposed to 8/6/4/2 penalty points. EBEL (1965, [29]) also reports increased reliability, using another variation (2 or 1 credit points, -2 or 0 penalty points, and 0.5 points if omitted).

GRITTEN & JOHNSON (1941, [38]) used Confidence Weighting for multiple-alternative testing. The candidates selected one answer (as in Pick-one Multiple Choice), and then indicated their degree of confidence in this choice on a 5-point scale. DRESSEL & SCHMID (1953, [28]) evaluated this method, using a 4-point confidence scale, and reported that the confidence-weighted scores proved slightly more reliable than those obtained by Pick-one MC. The validity was also estimated, using the results of a prior Multiple Choice test as criterion, and confidence-weighting scored slightly better than MC. However, the value of this criterion is questionable since it is itself an MC test; any score contamination caused by traditional MC scoring will be present in this criterion. Furthermore, the improvement was marginal; according to the authors "there is no evidence to indicate that the variation (..) can be attributed to anything but sampling fluctuation".

Using a short-answer test as a criterion, HOPKINS, HAKSTIAN AND HOPKINS (1973, [46]) found a slightly higher reliability but a slightly *lower* validity for Confidence Weighting (using a 3-point scale) when compared to conventional number-right MC.

Using an asymmetric 2-point scale (+2 or +1 if correct, –2 or 0 if incorrect), JENSEN (1983, [52]) found a distinctly higher reliability when compared to number-right MC scoring. He also reports a marked improvement in validity; in this case the criterion was the rating (by supervisors and personnel officers) of abilities at the workplace.

Comparing 'number-right minus number-wrong' to 'weighted-number-right minus weighted-number-wrong' (using a symmetric 2-point scale, +4/+1 or –4/–1, 0 points if omitted), BOKHORST (1986, [7]) again found higher reliability but lower validity for the weighted scores. His criterion was the overall achievement for the academic year.

HUNT (1982, [49]) investigated the effect of what he terms "Self-Assessment Responding" on learning. He concludes "It was found that the subjects in experimental treatments that required them to engage in SA responding learned the material in fewer trials than did subjects in the control group, who simply learned the material."

2.3 Probability Measurement

DE FINETTI (1965, [16]) proposed that we should consider the 'subjective probabilities' assigned by a student to *all* alternative answers for each item. As with Multiple Choice, each test item is presented with a set {1, 2, ..., **k**} of alternative answers. Instead of picking one, however, the student must respond by

assigning a personal probability r_j to each of the alternatives, where $0 \leq r_j \leq 1$ and $\Sigma r_j = 1$ for $j = 1, 2, ..., \mathbf{k}$. De Finetti discussed both theoretical and practical scoring rules which would take all probabilities r_j into account, presenting $s(i) = 2r_c - r_j^2$ as the most powerful.
However, SHUFORD, ALBERT AND MASSENGILL (1966, [70]) maintained that the item score should be based only on the probability r_c assigned to the correct answer. Based on extensive mathematical analysis, they concluded that only a logarithmic scoring rule would meet all their requirements. Since this is the basis for the Multiple Evaluation method used in our research, it will be dealt with in greater detail in the next chapter. It may be noted, in passing, that EBEL (1968, [31]) questioned the practical value of this system. He feared that the increased test and scoring times would outweigh the advantages.

A simple linear scoring rule was evaluated by MICHAEL (1968, [59]): 10 markers (each equivalent to 0.1 points) were to be distributed over all alternatives, and the item score was equal to the total points allocated to the correct answer. Compared to "number right" MC scoring, this method gave consistently higher reliability.

A few years later, RIPPEY (1970, [63]) compared five scoring rules:

a) linear, $s(i) = r_c$
(cfr. Michael);
b) truncated logarithmic, $s(i) = 1 + \frac{1}{2} {}^{10}\log(r_c)$ for $r_c \geq 0.01$, else $s(i) = 0$
(cfr. Shuford);
c) spherical, $s(i) = r_c / \sqrt{}(\Sigma r_j^2)$;
d) euclidian, $s(i) = 1 - \sqrt{}(\Sigma(r_j - k_j)^2 / 2)$ where k_j is the mean probability assigned to the j-th option by a criterion group of experts;
e) inferred choice, $s(i) = 1$ if $r_c > r_j$ for all $j \neq c$, otherwise $s(i) = 0$
(cfr. Pick-one MC).

The score reliabilities were highest for the linear rule (a) and lowest for inferred choice (e). The logarithmic (b) and spherical (c) rules proved little better than inferred choice.
KOEHLER (1974, [54]) compared de Finetti's quadratic rule, Shuford's logarithmic rule, Rippey's spherical rule, number-right MC scoring and MC with correction for

guessing. Finding little difference in reliability between these scores and a greater correlation between Probability Measurement and a measure of “overconfidence”, he concluded that the conventional number-right procedure is to be recommended. Further doubts on probability measurement were cast by SIEBER (1974, [77]) when she found that the confidence reported by the students was influenced by the importance of the test results. HAKSTIAN AND KANSUP (1975, [39]) compared the linear rule to conventional number-right; they found a higher reliability for the linear rule, but a lower validity when the scores were correlated with the semester-end grades.

Distinctly higher reliabilities for Probability Measurement scoring systems are again reported by PUGH AND BRUNZA (1975, [61], ‘linear’ rule), who also checked for personality bias but found no significant correlations; a few years later by POIZNER, NICEWANDER AND GETTYS (1978, [60], logarithmic rule), who deliberately manipulated the subject’s knowledge in a controlled experiment; and again by ABEDI AND BRUNO (1989, [1], logarithmic rule), using a simplified pencil&paper version for 3-alternative items.

FRIEDLAND AND MICHAEL (1987, [36]) evaluated the reliability and validity of probability measurement scoring, using the rating by supervisors on eight job performance dimensions and eight factor scales of standardized personality inventory as a criterion. Comparing ‘number right’, ‘linear probability’ and ‘formula probability’ scores, they found only marginal differences in reliability and validity.

2.4 Other methods

Free choice. Examinees mark as many answers in an MC test as they consider possibly correct. If the correct answer is included, credit points are awarded: the fewer alternatives marked, the higher the credit. On the other hand, if the correct answer is not included a penalty is calculated: the more (incorrect) alternatives marked, the more severe.

DRESSEL & SCHMID (1953, [28]) compared this system to both Pick-one MC and Confidence Weighting, and found that it gave the worst results both for reliability and for validity.

Elimination. COOMBS (1953, [12]) proposed a procedure where students must cross out all alternatives that they consider incorrect. They would score 1 point for each wrong answer eliminated, but if they crossed out the correct answer they received a penalty of 1–**k** points.
A few years later, COOMBS, MILHOLLAND AND WOMER (1956, [13]) compared this system to traditional number-right MC scores, and reported a higher reliability for the 'elimination' scores.
ARCHER (1962, [4]) compared 'free choice', 'elimination' and conventional MC testing. Although the reliability of MC was slightly lower than the other two measures, the validity of MC proved slightly higher (compared to the criterion: the student's rank in the group).
Furthermore, both of these systems are only applicable for *multiple*-choice items with a sufficient number of distractors to allow for score discrimination. In the trivial case where **k** = 2, crossing out one alternative automatically implies that the other must be correct. The studies mentioned above used **k** = 4 items.

Graded distractors. Instead of 'all or nothing' (full points for the correct answer and zero for all distractors), some points may be awarded to one or more of the distractors. For example: (a) is the correct answer and scores 10 points, (b) is close and scores 5 points, (c) and (d) are totally wrong and score zero points (or maybe even penalty points, for serious fallacies). A more sophisticated version uses the euclidian scoring rule (RIPPEY, [63]), and involves the pre-scoring of all items by a criterion group consisting of experts.
In most cases this system would appear to be of dubious value since, in effect, it considers some incorrect answers to be 'less wrong' than others. This may be defensible in some specific applications, e.g. when testing reading skills, but it is not suited to our (technical) environment where answers to examination questions are usually either 'right' or 'wrong'.

2.5 Tabular overview

A brief summary of earlier research is given in Tables 1 and 2. The scoring methods described and/or evaluated are indicated as follows:

- Pick One:
  - P1 = number right = pick one = conventional MC;
  - P1C = P1 with correction for guessing,
    item score = number_right – number_wrong/(**k**-1).
- Confidence:
  - CW = Confidence Weighting: if right e.g. +3/+2/+1, if wrong 0;
  - CWC = CW with correction for guessing (weighted penalty if wrong).
- Probability:
  - MEL = Multiple Evaluation, Logarithmic scoring rule (cfr. Shuford);
  - PMO = Probability Measurement, Other rules (linear, quadratic, etc.).

Results for reliability (r) and validity (v) are shown as – = reject, o = neutral, + = recommend. Table 1 summarizes the results of comparative studies, and Table 2 gives the results when only one of the scoring systems was evaluated.
Figure 4 illustrates the numerical results for reliability comparisons, insofar as the relevant data is available. However, it should be noted that the test reliability is calculated in various ways by different researchers, so these results should only be seen as an indication of the relative performance of the scoring systems.

Although these surveys are necessarily brief and sketchy, they may serve to illustrate a general trend. ‘Pick One’, without or with Correction-for-guessing, is rejected almost unanimously. ‘Confidence weighting’ is considered more favourably, except in the studies that also considered ‘Probability Measurement’. The latter appears to be the clear winner, certainly for test reliability.
In all fairness, however, it should be noted that many authors express their doubts concerning the test validity when Probability Measurement is used: they fear that the results may be contaminated by various ‘personality factors’. We will return to this aspect later.

Reference	Pick One		Confidence		Probability	
	P1	P1C	CW	CWC	PMO	MEL
Frary, Cross, Lowry [35] 1977	o	–				
Leclerq [55] 1983	–	–				
Hughes [48] 1989	–	–				
Hevner [45] 1932	–	–	r+v+	–		
Soderquist [79] 1936		–		r+		
Swineford [80] 1938		–	r+			
Gritten, Johnson [38] 1941	o	o	o	o		
Dressel, Schmid [28] 1953	–			r+ v+		
Ebel [29] 1965	–			r+		
Ahlgren [2] 1969	r–		r+			
Hopkins, Hakstian, Hopkins [46] 1973	r– v+			r+ v–		
Hunt [49] 1982	–		+			
Jensen [52] 1983	–			r+ v+		
Bokhorst [7] 1986		r– v+		r+ v–		
Hunt [50] 1993	–		+			
Shuford [71] 1967	–					r+
Michael [59] 1968	–				r+	
Rippey [63] 1970	r–				r+	r–
Koehler [54] 1974	+	o			o	o
Hakstian, Kansup [39] 1975	r– v+				r+ v–	
Pugh, Brunza [61] 1975	–				r+	
Poizner, Nicewander, Gettys [60] 1978	–				+	++
Rippey, Voytovich [64] 1983	–					+
Friedland, Michael [36] 1987	o				o	
Abedi, Bruno [1] 1989	–					r+
Hambleton, Roberts, Traub [40] 1970	r+ v–		o			r– v+
Shuford, Brown [73] 1975	–		–			+
Bruno [11] 1993	–	–			r+	

Table 1. Comparisons of scoring systems. 'Confidence weighting' and 'Probability measurement' lead to higher reliabilities in most studies, but the results for validity vary and many authors warn that other (personality) factors may contaminate the results.

Reference	Pick One		Confidence		Probability	
	P1	P1C	CW	CWC	PMO	MEL
Trow [84] 1923		–				
Ziller [89] 1957		–				
Slakter [78] 1968		–				
Harris [42] 1969		–				
Bachman, Palmer [5] 1996		–				
Tarone, Yule [82] 1989			+			
Hassmén, Hunt [43] 1994			+			
de Finetti [16] 1965					+	
Shuford, Albert, Massengill [70] 1966						+
Ebel [31] 1968						–
Baker [6] 1968						+
Shuford [72] 1969						+
Gardner [37] 1969						+
Sibley [76] 1974						+
Sieber [77] 1974						–
Shuford [75] 1993						+
Dirkzwager [21] 1996						+

Table 2. Evaluations of scoring systems. Ebel (1968) reviewed Shuford's 'ScoRule' and concluded that it was too complicated for practical use. Sieber (1974) found that a personality factor (the perceived importance of the test) influenced the scores.

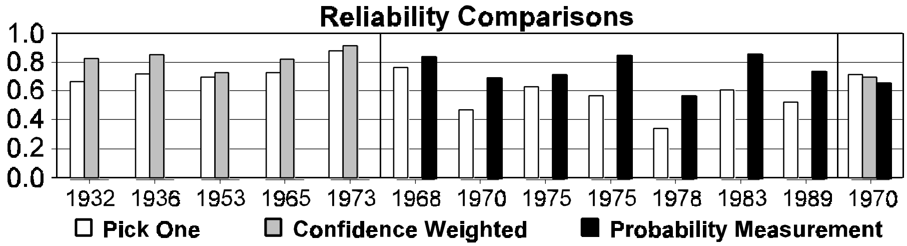


Figure 4. Numeric comparisons of reliability from empirical research. Note that test reliabilities may be calculated in different ways.

2.6 Summary, tentative conclusions

Conventional Multiple Choice ("Pick One") scores contain an 'unknowable' guessing effect, and MC with correction for guessing has been (consistently) rejected since 1923. Even so, P1 or P1C is still often considered "the best of a bad lot", based on:

- efficiency, when other (more complex) scoring systems involve greater test and administration time;
- reliability, when other methods produce similar or only marginally better results (so that 'efficiency' and/or 'validity' results may lead to rejection of confidence-derived scores in favour of MC);
- validity, when 'confidence' results seem contaminated by other factors.

The 'efficiency' problem associated with test and administration time for confidence testing could indeed offset the advantages of reduced test length due to greater reliability. However, when using computers for testing and evaluation this problem tends to disappear.

The 'reliability' of confidence- or probability-derived scores is never less than conventional MC, and often found to be significantly better. The worst case results (often for highly simplified versions of confidence testing) are similar or 'show no significant difference'.

On the 'validity' aspect, the jury is still out. When rated on the basis of correlation with some criterion, some results are better and others are worse. Note that several 'confidence weighting' methods were compared, and several different 'criteria' were used:

• CW better than MC:	Hevner [45]; criterion: general background, Dressel et al. [28]; criterion: prior MC test, Jensen [52]; criterion: subjective rating.
• inconclusive result:	Friedland et al. [36]; criterion: subjective rating.
• MC better than CW:	Hopkins et al. [46]; criterion: short answer test, Hakstian et al. [39]; criterion: semester-end grades, Bokhorst [7]; criterion: semester-end grades.

Of these researchers, only Hakstian used a (simple, linear) version of Multiple Evaluation; all others used a Confidence Weighting system.

Understandably, a major cause for concern is the possibility that 'Confidence Weighted' or 'Probability Measurement' test results may be contaminated by personality traits such as lack of confidence or a tendency towards risk-taking. This undesirable effect is often cited, sometimes theoretically and sometimes empirically. However, several authors claim that the effect tends to disappear after some practice:

- significant effect found:
  Jacobs [51]: CW reliability influenced by credit-penalty relationship,
  Hansen [41]: response style depends on risk-taking or risk-avoiding tendency,
  Koehler [54]: PM more strongly correlated with confidence in general,
  Sieber [77]: PM scores influenced by perceived importance of test.
- effect tends to disappear after practice:
  Shuford et al. [70],
  Echternacht et al. [33]: no correlation with twelve personality variables,
  Sibley [76]: most subjects are quite realistic after only a few tests,
  Pugh et al. [61]: no correlation with risk taking vs. caution measures,
  Leclercq [56]: continuous improvement over three consecutive tests,
  Dirkzwager [21]: eleven year olds soon learn to give realistic estimates.

To summarise:

- Confidence- or probability-derived scores appear to offer greater test reliability than conventional Multiple Choice.
- Provided computers are used, the efficiency should be at least equivalent to that of conventional tests; possibly even better, if shorter tests prove sufficient.
- Special attention must be paid to the test validity. However, it is claimed that practice (repeated tests) will eliminate the undesirable effects of personality factors.

3 - Theoretical background of Multiple Evaluation

In the 1950's and early 1960's several researchers (notably de Finetti [15], van Naerssen [86], Toda [83], and Roby [66]) were experimenting with test procedures involving probability measurement. In an article published in 1966 [70] Shuford, Albert and Massengill integrate and generalize the then-known results. They first define the type of scoring system that is permissible in an educational environment, as follows:

> “Admissible probability measurement procedures utilize scoring systems with a very special property that guarantees that any student, at whatever level of knowledge or skill, can maximize his expected score if and only if he honestly reflects his degree-of-belief probabilities.” *(p.125)*

3.1 Derivation of the logarithmic scoring rule

In the first section of Shuford’s 1966 paper [70], dealing with those instances in which the possible answers to a question are stated in the test itself, the above definition is refined in three steps.

> “Consider one of the items in such a test and let the set of possible answers be associated with the set {1, 2, ..., **k**}. The student (..) responds by assigning a probability r_1 to the event that alternative 1 is the correct answer, r_2 that 2 is the correct answer, and so on, where
>
> $0 \le r_j \le 1$ and $\Sigma r_j = 1$ ($j = 1, 2, ...,$ **k**)
>
> The student receives a score according to a scoring system, $f_j(r_1 \ldots r_k)$. If (..) the correct answer is, in fact, c, then the student receives a score in the amount $f_c(r_1 \ldots r_k)$.
>
> Let the student’s uncertainty as to the correct answer be represented by a probability distribution, $\bar{p} = (p_1, p_2, ..., p_k)$, over the set {1, 2, ..., **k**}, where $0 \le p_j$ and $\Sigma p_j = 1$. Then the student’s expected score, should he make the probability assignments $r_1, r_2, \ldots r_k$, is
>
> $E[S(\bar{r},\bar{p})] = \Sigma p_j \times f_j(\bar{r})$ ($j = 1, 2, ...,$ **k**)

where, for notational convenience, we let $\bar{r} = (r_1, r_2, ..., r_k)$ and $\bar{p} = (p_1, p_2, ..., p_k)$.

If the student is assumed to have knowledge of $f_j(\bar{r})$ and, of course, $\bar{p}$ and desires the highest possible expected total test score, he should be advised to respond with a probability assignment (i.e. choose $\bar{r}$) so as to maximize $E[S(\bar{r},\bar{p})]$ *(..)*. Now consider the following question: under what conditions on the scoring system, $f_j(\bar{r})$, is it always in the best interest of the student to respond with the probability assignment $\bar{r} = \bar{p}$? Scoring systems which satisfy these conditions will be called *reproducing scoring systems* (or RSS's for short).

(..)

Definition: The scoring system $f_j(\bar{r})$ is said to be a RSS if and only if the expected score $E[S(\bar{r}, \bar{p})] = \Sigma p_j \times f_j(\bar{r})$ (j = 1..**k**) is maximized when $E[S(\bar{r}, \bar{p})] = E[S(\bar{p}, \bar{p})]$ for all allowed values of r_j and p_j." *(p. 127-128)*

The class of RSS's is virtually inexhaustible *(p.129)*. In the same paper, 'Theorem 2' explains how any bounded nonnegative function h(t) with h(1) = 0 and a bounded derivative for 0 ≤ t ≤ 1 can be used as the basis for constructing a (bounded, continuously differentiable) RSS for **k**=2. The scoring rules for the two alternative answers (e.g. True and False) are then derived from h(t) as:

$$s_1(r) = \int_0^r h(t)dt \quad \text{and} \quad s_2(r) = \int_{1-r}^{1} \frac{t}{1-t} h(t)dt$$

As an example, we can take h(t) = 1−t. This "bounded, non-negative function with h(1) = 0 and a bounded derivative for 0 ≤ t ≤ 1" leads to the symmetrical RSS scoring rule $s_{1,2} = r_c - ½r_c^2$.

We can also take h(t) = 1−t², which meets the same criteria. However, this function results in a set of scoring rules, one rule for each of the two alternative answers: $s_1 = -⅓r_c^3 + r_c$ and $s_2 = ⅓r_c^3 - 1½r_c^2 + 2r_c$. In this case the item score depends on which of the two alternatives is correct; the maximum score for option 1 is 2/3, whereas for option 2 it is 5/6. Note that r_c is the response to the correct alternative; for a given response r, $r_1 = r$ and $r_2 = (1–r)$.

These two examples are illustrated in Figure 5.

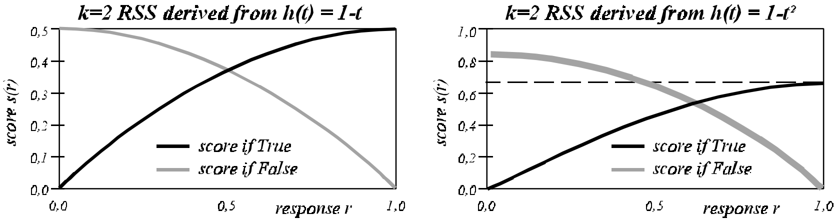


*Figure 5. Two examples of RSS rules for **k** = 2, derived according to Shuford's Theorem 2.*

As Shuford points out:

> "The construction of a RSS (..) allows a certain asymmetry to exist between the scores received by the student depending on which of the answers was correct. Under most circumstances, it probably would be desirable to arrange things so that the score received by a student who assigns a probability *r* to the correctness of an answer arbitrarily designated 1 is the same as the score he would receive if he assigned the same probability *r* to the correctness of the same answer arbitrarily designated 2." *(p.130-131)*

A RSS that meets this requirement is called a *symmetric reproducing scoring system* (SRSS). As proved by Shuford *et al.* in their paper, both the so-called "spherical" and "quadratic" scoring systems satisfy all requirements outlined so far:

> "For the spherical, we can take
>
> $s(i) = r_c \,/\, (\Sigma r_j^2)^{½}$
>
> and for the quadratic, we can choose
>
> $s(i) = 1 + r_c^2 - (1 - r_c)^2 - \Sigma r_j^2$ " *(p.135)*

Both of these systems have a serious drawback, however. The score is not only determined by the probability assigned to the correct answer, but also by the way in which the student's uncertainty is distributed over the other alternatives. For dichotomic items (True/False) this is quite permissible, since if a probability r_c is assigned to the correct answer it follows that a probability $(1-r_c)$ must be assigned to the incorrect alternative.

For items with three or more alternatives we can only state that the sum of the probabilities assigned to the other alternatives must be $(1-r_c)$, and when using the quadratic or spherical scoring systems the item score will depend on the distribution of this probability assignment over the incorrect alternatives. To be precise: the item score will be minimized when the total remaining probability $(1-r_c)$ is assigned to *any* one of the incorrect alternatives and the probability assignment for all other incorrect alternatives is zero. Figure 6 illustrates this effect for **k** = 3.

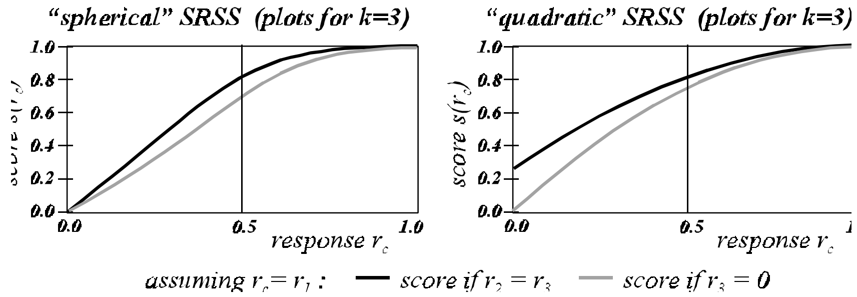


Figure 6. *The so-called "spherical" and "quadratic" SRSS's have a serious drawback: for **k** > 2 the item score depends on the distribution of remaining probability over the distractors. The maximum item score is obtained when all distractors seem equally (un-)likely; eliminating some but not all of the distractors will reduce the item score.*

In most cases, this dependence will be undesirable - even impractical, since taking the responses for all incorrect alternatives into account imposes rigorous restraints on item construction. It also runs contrary to the results obtained by earlier researchers using Coombs' elimination scoring system: for the spherical and quadratic rules, the more distractors a student can eliminate, the *lower* the final score will be!

Therefore:

> "A question of great practical interest revolves about the possibility of constructing a RSS (symmetric or otherwise) in which the student's score depends only on the probability assigned to the correct answer – not on the probabilities assigned to the other, incorrect answers." *(p.135)*

The paper goes on to prove that:

> "... the "logarithmic" scoring system is the only one which has the property that the student's score depends only on the probability that he assigns to the correct answer when there are more than two possible answers. All other RSS's lack this property." *(p.136-137)*

DIRKZWAGER [22] derives the same result as follows.

The item score, *s(i)*, is based on the response, *r(i,j)*; in fact it should reflect the response $r_c(i)$, the probability assigned to the correct alternative. We also wish to ensure that the response is based on a realistic estimate of the true (personal) probability that this is indeed the correct answer: for each item, *r(j)* should reflect the true probability *p(j)* that alternative *j* is the correct answer for this item.

To this end, the expected item score E(s) = $\sum p(j)\times f(r(j))$ must be maximized when $r(j) = p(j)$ for all values of *j*. We can determine the partial derivative of this function, under the boundary condition that $\sum r(j) = 1$, with the aid of the Lagrange multiplier λ :

$$\frac{\delta(\Sigma[p(j)\times f(r(j))] + \lambda\times[1-\Sigma r(j)])}{\delta r(j)} = 0 \quad \text{for all values of } j$$

$$\Rightarrow \quad p(j)\times\frac{\mathrm{d}f(r(j))}{\mathrm{d}r(j)} - \lambda = 0$$

$$\frac{\mathrm{d}f(r(j))}{\mathrm{d}r(j)} = \frac{\lambda}{p(j)} \qquad \text{with } p(j) = r(j) > 0 \text{ for all values of } j \text{, so}$$

$$\frac{\mathrm{d}f(r(j))}{\mathrm{d}r(j)} = \frac{\lambda}{r(j)} \qquad \text{where } \lambda \text{ is a constant, independent of } r(j)$$

$$\Rightarrow \quad f(r(j)) = \mathrm{A}\times\ln(r(j)) + \mathrm{B} \qquad \text{with A and B as constants}$$

We may choose A and B such that the item score *s(i)* = 0 when *r(i,j)* = 1/**k** for all *j* (all alternatives considered equally probable) and *s(i)* = 1 when $r_c(i)$ = 1 (maximum probability assigned to the correct alternative for this item, so that *r(i,j)* = 0 for all incorrect alternatives).

The item score *s(i)* is then determined by $r_c(i)$ as follows:

$$s(i) = f(r_c(i)) = \frac{\ln(r_c(i))}{\ln \mathbf{k}} + 1 \qquad (3.1)$$

In this basic format, the logarithmic scoring rule is too severe. To quote Shuford *et al.*:

> "We find, however, that the logarithmic scoring system is unbounded and thus impossible to realize in practice: how can one give a student a score of minus infinity?" *(p.137)*

As a possible solution, they suggest that we truncate the logarithmic scoring function at some small value of $r_c > 0$. For example:

$$\textbf{item score } s(r_c) = \begin{Bmatrix} 1 + {}^{10}\log r_c & \text{for } .01 < r_c \le 1 \\ -1 & \text{for } 0 \le r_c \le .01 \end{Bmatrix} \qquad (3.2)$$

This truncated system has the reproducing property for $.027 \le p_c \le .973$. For extreme values ($p_c \le .027$ or $p_c \ge .973$) there is some loss of accuracy: the student's score is maximized in these ranges by setting $r_c = 0$ or $r_c = 1$, respectively. It should be noted that this rule may appear relatively lenient, with a maximum score and maximum penalty of +1 and –1 respectively, but this is deceptive. The 'zero-knowledge' score for $r_c = 1/\mathbf{k}$, s = 1 + log(1/**k**), is greater than 0. Relative to this 'zero' score the maximum penalty is greater than the maximum score and depends on the number of options: the ratio between (maximum) penalty and score is 1.86 for **k** = 5, increasing to 5.64 for **k** = 2.

DIRKZWAGER [22] provides a more elegant solution.

We can introduce a *tolerance parameter* (**t**), chosen in the range 0 ≤ **t** < 1/**k**.

The probability assignment *r(i,j)* is then converted to a more modest probability assignment *r'(i,j)* = (1 − **t.k**)×*r(i,j)* + **t**, so that:

$r(j) = 0 \Rightarrow r'(j) = \mathbf{t}$

$r(j) = 1/\mathbf{k} \Rightarrow r'(j) = 1/\mathbf{k}$ (unchanged for *r(i,j)* = 1/**k** : *s'(i)*=*s(i)*=0)

$r(j) = 1 \Rightarrow r'(j) = 1 - \mathbf{t.k} + \mathbf{t}$

In effect, this conversion limits the response to the range $\mathbf{t} \le r'(j) \le 1 - (\mathbf{k}-1)\times\mathbf{t}$ and presents this to the student on a scale $0 \le r(j) \le 1$. We are taking the student's probability assignments "with a pinch of salt"; most importantly, we are taking "impossible" (r = 0%) as meaning "highly unlikely" (r'= t), thus limiting the worst-case negative score to a finite value.

In the strictest sense, therefore, this is no longer a RSS: the item score is not maximized when the probability assignment $r(j)$ is exactly equal to the true probability $p(j)$. For practical purposes this is quite acceptable, however, especially for small values of **t**. All other advantages of the logarithmic scoring system are maintained, without the loss of smoothness introduced by truncating the scale as suggested by Shuford.

This conversion also effects the maximum achievable score, since $r_c = 1$ is taken to mean $r_c' \ge 1-(\mathbf{k}-1)\mathbf{t}$:

$$s'_{i,\,\max} = \frac{\ln(1-\mathbf{t}.\mathbf{k}+\mathbf{t}) + \ln(\mathbf{k})}{\ln(\mathbf{k})} < 1 \quad (\text{for } \mathbf{t} > 0)$$

To compensate for this we multiply the item score by $1/s'(i)_{\max}$, and obtain the following scoring rule:

$$\boxed{s(r_c(i)) = \frac{\ln[(1-\mathbf{t}.\mathbf{k})\times r_c(i) + \mathbf{t}] + \ln(\mathbf{k})}{\ln(1-\mathbf{t}.\mathbf{k}+\mathbf{t}) + \ln(\mathbf{k})}} \tag{3.3}$$

(with $0 \le t < 1/\mathbf{k}$ as tolerance parameter)

If we select zero tolerance ($\mathbf{t}$ = 0) this rule reverts to the basic logarithmic scoring system given earlier, (3.1), with a maximum item score $s(1) = 1$ and a minimum score (maximum 'penalty') of minus infinity.

On the other hand, the maximum value for the tolerance parameter (with **t** approaching 1/**k**) yields a minimum score $s(0)$ approaching $1/(1-\mathbf{k})$, as can be derived by l'Hôpital's rule:

$$\begin{aligned}
\lim_{\mathbf{t}\to 1/\mathbf{k}} s(0) &= \left.\frac{\mathrm{d}(\text{numerator})/\mathrm{d}\mathbf{t}}{\mathrm{d}(\text{denominator})/\mathrm{d}\mathbf{t}}\right|_{r_c = 0;\ \mathbf{t} = 1/\mathbf{k}} = \\
&= \left.\frac{(1-\mathbf{k}.r_c)/[(1-\mathbf{t}.\mathbf{k})r_c+\mathbf{t}]}{(1-\mathbf{k})/(1-\mathbf{t}.\mathbf{k}+\mathbf{t})}\right|_{r_c = 0;\ \mathbf{t} = 1/\mathbf{k}} = \\
&= \left.\frac{1/\mathbf{t}}{(1-\mathbf{k})/(1-\mathbf{t}.\mathbf{k}+\mathbf{t})}\right|_{\mathbf{t} = 1/\mathbf{k}} = \\
&= \frac{\mathbf{k}}{(1-\mathbf{k})/(1/\mathbf{k})} = \frac{1}{(1-\mathbf{k})}
\end{aligned}$$

In this case, pure guesswork (selecting $r(j) = 1$ for one of the alternatives at random) will yield an expected item score $E[s(i)] = (1/\mathbf{k})\times s(1) + ((\mathbf{k}-1)/\mathbf{k})\times s(0) = (1/\mathbf{k}) - (1/\mathbf{k}) = 0$. This effectively degrades the scoring rule to a variation on Multiple Choice with traditional correction for guessing, where the expected score is maximised by setting $r_j = 1$ when p_j is considered largest of all p_k for this item.

3.2 Early experiments

Although the theoretical basis has been known for at least thirty-five years, relatively little practical research is reported. Many of the earliest experiments involved paper&pencil testing with look-up tables for scoring according to the logarithmic rule. Looking back on this period in a Keynote Address in 1993 [75], SHUFORD prefaces a brief explanation of his 'SCoRule' system with the words:

> "I undertook a long pilgrimage in search of a way to implement this procedure without using a computer. I failed - figures 2, 3 and 4 from that pilgrimage may help you understand one reason why we need a computer."

These illustrations are reproduced here in Figure 6 as A, B and C.

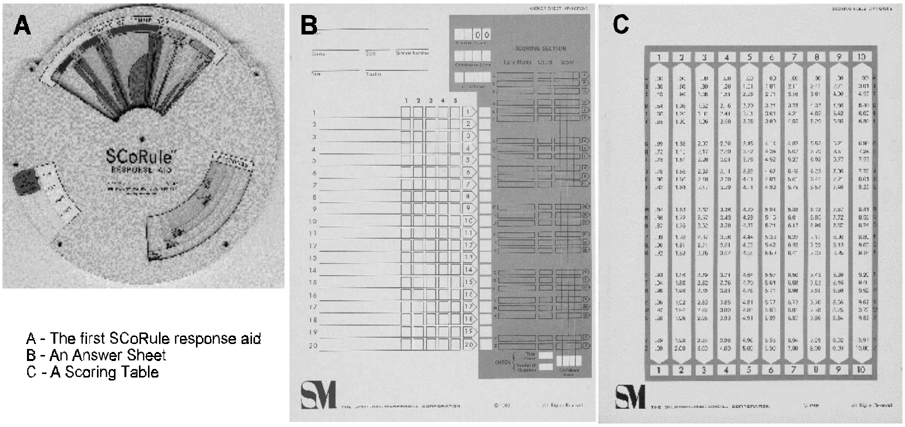


Figure 6. Shuford's first SCoRule. (Reproduced from Shuford, 1993.)

Using the disc A (which displayed the relationship between probability and score) the student would spread a total probability of one over the possible answers. He then copied letters from the response area into corresponding boxes on his answer sheet B. Finally, the letter codes were tallied and the table C was used to compute the total score for the test.

It is not surprising that EBEL (1968b, [31]) questioned the practical value of this system. Although the higher reliability could lead to less items per test, the increased test and scoring times due to the complexity of the task could easily outweigh this advantage.

However, even this rather complicated test environment gave promising results, witness GARDNER's report [37] in 1969:

> "The Academic Instructor and Allied Officer School (AIAOS) has been experimenting with confidence testing techniques since July 1968. (...) Although research is still continuing along this line, the data that have been gained and analyzed indicate that confidence testing has great potential as a diagnostic tool."

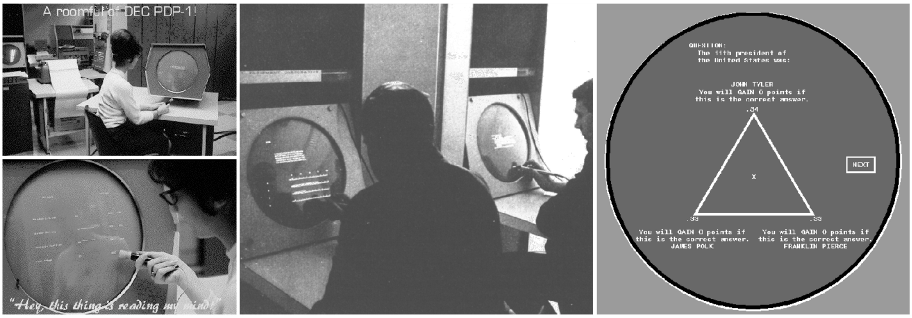
(Note that the reference to 'confidence testing' may be misleading: the method used in this research was the paper&pencil version of the logarithmic rule, as described above.)

In following years, early experiments with computer-aided testing are reported (e.g. BAKER [6] in 1968, SIBLEY [76] in 1974 and DIRKZWAGER [19] in 1975). Given the state of the art at that time, this was hardly a viable proposition for most educational institutes. As an example, Sibley describes his test environment as follows:

> "As you will see from the figures that follow, the computer-based system we have developed requires a computer terminal with graphics capabilities and some means of allowing the student to interact with it in elapsed times of at most a few seconds. The original system was developed on a highly interactive console connected to a powerful IBM 370/155 computer. A newer system exists on a 'smart terminal', the IMLAC PDS-1D, which supports the system independently of any larger computer."

Even using the latter system, the user interface (screen display) and feedback look primitive by modern standards - as illustrated in Figure 7. The results (with 66 test subjects of which 15 took more than one test) were sufficiently promising, however, to warrant hopes for the future:

> "... An implementation of our ideas on the University of Illinois' PLATO IV system..."


A roomful of DEC PDP-1!
"Hey, this thing's reading my mind!"
QUESTION:
The 11th president of
the United States was:
JOHN TYLER
You will GAIN 0 points if
this is the correct answer.
.34
NEXT
X
.33
.33
You will GAIN 0 points if
thie is the correct anewer.
JAMES POLK
You will GAIN 0 points if
thie ie the correct anewer.
FRANKLIN PIERCE

Reproduced from Shuford (1965, 1993), *Cybernetic Testing*

Reproduced from Baker (1968), *The uncertain student and the understanding computer*

Based on Sibley (1974), *Computer Assisted Admissible Probability Testing* [redrawn]

Figure 7. Some early experiments with computer-aided probability testing.

It wasn't until the late 1980's that computers became available (and affordable) in sufficient quantities in practical educational environments. RIPPEY (1986, [65]) was one of the first, using an Apple II microcomputer. DIRKZWAGER's TestMe (1993, [20]) was available for Macintosh, Atari and MS-DOS machines; it was subsequently upgraded to an MS-Windows version, TestBet (1996-2001, [26]). The latter program was used in our research, for computer-aided testing with **k**=4 items.

Furthermore, most research (even in recent years) was designed to test and expand the theory. In particular, researchers were mainly interested in 'realism' and 'test reliability' (two essential prerequisites, admittedly) and not so much on the adaptations required if this method is to be applied for certifying assessment in a regular educational environment. This can be illustrated with just two recent examples.

DIRKZWAGER [21] reports in 1996:

> "Testing with personal probabilities is an advantageous alternative to multiple choice testing because guessing is eliminated. A crucial condition is that the subjects should be well calibrated. They are well calibrated when they report their true probabilities of being right. It is shown to be feasible in an ordinary school setting to have even 11 year olds learn to be well calibrated in a short time, especially if probability testing is the method they become accustomed to. An interactive computer program to take a test (TestBet), that was used in the experiment, is described. Its use in regular education is feasible, and the educational advantages are discussed."

The bulk of this report is devoted to 'realism' measurements and the learning curves of the pupils involved. Almost as an afterthought, Dirkzwager notes in his discussion:

> "From an educational point of view we may state that by this testing method people learn to use and trust the knowledge they have, however little it may be."

As far as ‘educational advantages’ are concerned, the emphasis is on the diagnostic and educative effects, not on the accuracy and reliability as an assessment tool.

A later paper by the same author [23] in 1998 bears the promising title: *‘A comparison of personal probability scoring, certainty, confidence scoring, Rasch estimates, Multiple choice scoring and realism measured with TestBet’.* The subjects in this case were ‘two parallel groups (n = 32 and n = 23) of about 14 year old High School students’. Two short quotes from the ‘Conclusions’:

> “It is shown that Multiple Evaluation (..) is a more reliable and valid measurement method than Multiple Choice or its Rasch extension.
>
> (..)
>
> It seems also crucial that the subjects have some experience with the method: the results at the second session were an improvement upon the first session and the full profitability of Multiple Evaluation as implemented in TestBet will only show when the method is as common to the subjects as Multiple choice is now.”

As such, this is a valuable report in that it tends to conclusively prove the reliability of ME scoring as opposed to other methods. However, when looking at ‘scores’ that range from –1230.68 to +1857.40 it is difficult to envisage the consequences on a ‘standard’ scale of 1 to 10 or A to F.

3.3 ME: a tool for certifying assessment?

We do not wish to in any way underrate the value of ME as a ‘learning tool’, a method to teach students ‘realism’ and train them to ‘use their (possibly limited) knowledge effectively’. These are valuable assets, in any educational environment.
However, we would like to utilise the advantages of this scoring system for certifying (‘pass’/’fail’) tests. Given that suitable computers are now more readily available and that Multiple Evaluation apparently gives more reliable results than traditional methods, it should be possible to obtain more reliable certifying scores in this way. By this we mean scores on a standard scale, comparable to current

scores, but with the increased accuracy that ME appears to provide. To go one step further: since current tests are already reasonably accurate in our opinion and experience, how much testing time can we save by using a system that provides at least equally accurate results in a shorter time?

As it stands, however, the logarithmic scoring rule leads to scores that are difficult to interpret and not compatible with standard scoring ranges. Furthermore, it is unclear which point on the scale can be considered a cutting score - i.e. the borderline between pass and fail.

Some further development is required, and our research proposes solutions to practical problems and evaluates the results in a regular educational environment.

4 - Implementation of ME scoring

On the basis of earlier work, the hypothesis is that Multiple Evaluation should prove a valuable test method and a major improvement over Multiple Choice. Furthermore both Shuford and Dirkzwager prove that the logarithmic scoring rule is the *only* proper scoring rule for items with more than two alternatives, if the item score is to be based exclusively on the response for the correct alternative. The problems associated with worst-case scores of minus infinity can be resolved by means of the tolerance parameter suggested by Dirkzwager.
Before actually implementing this system, however, some further practical points must be resolved.

4.1 Choice of tolerance parameter

Chapter 3 presented a modified logarithmic scoring rule (3.3) which incorporates a tolerance parameter **t**, where $0 \leq \mathbf{t} < 1/\mathbf{k}$. As we have seen, selecting $\mathbf{t} = 0$ makes the rule too severe, whereas when **t** approaches 1/**k** the rule becomes too tolerant.

Between these two extremes, the tolerance parameter can be selected to achieve any desired minimum item score $s(0)$. This means that we can set the ratio between the minimum and maximum achievable item scores, thereby determining how many items must be answered perfectly (giving $s(1) = 1$) to compensate for one maximally incorrect item (with a probability assignment $r_c(i) = 0$ for the correct alternative).
If one maximally incorrect item is to be compensated by **T** maximally correct items:

$$s(0) + \mathbf{T} \times s(1) = 0 \;\Rightarrow\; \mathbf{T} = -s(0)/s(1) = -\frac{\ln(\mathbf{t}) + \ln(\mathbf{k})}{\ln(1 - \mathbf{tk} + \mathbf{t}) + \ln(\mathbf{k})}$$

If we select $\mathbf{T} = 1$ for example, i.e. $s(0) = -s(1)$, it can be shown that $\mathbf{t} = 1/(\mathbf{k} \times (\mathbf{k} - 1))$ where **k** is the number of alternatives. (For $\mathbf{k} = 2$ this would result in $\mathbf{t} = 0.5$, but given the restriction $\mathbf{t} < 1/\mathbf{k}$ this must be limited to $\mathbf{t} = 0.4999...$)

For several practical values of **T** and **k**, the corresponding values of **t**(**T**,**k**) were derived by numeric approximation. The results are given in Table 3.

	k=2	k=3	k=4	k=5
T=1	0.4999	0.1667	0.0833	0.0500
T=2	0.1910	0.0447	0.0174	0.0086
T=3	0.0804	0.0134	0.0041	0.0016
T=4	0.0362	0.0043	0.0010	0.0003

*Table 3. The tolerance parameter **t(T,k)** for several values of the number of alternatives (**k**) and tolerance ratio **T**.*

The effect of these values for **t** is clearly illustrated in Figure 8. For **k** = 2, .., 5 the ME item scores *s(i)* are plotted as a function of the probability $r_c(i)$ assigned to the correct alternative, for the four corresponding values of **t** given in Table 3.
It is apparent that the ME scoring rule becomes progressively more severe for very low probability assignments $r_c(i)$ as the tolerance **t** is decreased, whilst the effect for 'correct' answers ($r_c(i)$ > 1/**k**) is negligible. The tolerance parameter can therefore be seen as a means to limit the penalty points for students who consider the *correct* item to be highly *unlikely*. In that sense it sets the tolerance for (random) guessing and serious misconceptions. This is especially true for **k** > 2, where the effect only becomes clearly noticeable for probability assignments $r_c(i)$ < 5%.

On the other hand, introduction of the tolerance parameter results in some deviation from the requirement for a true RSS. The score is no longer maximized when *r(i)* = *p(i)*, since *r(i)* is converted to some value *r'(i)* that is closer to 1/**k** and the maximum score is obtained when this new value *r'(i)* equals *p(i)*. The deviation is proportionate to the tolerance **t**, and is a function of the response: *dev* = |*p(i)* – *r(i)*| = |**t** – **t**×**k**×*r(i)*|. This is zero for *r(i)* = 1/**k**, increasing linearly to **t** for *r(i)* = 0 and to (**k**–1)×**t** for *r(i)* = 1.

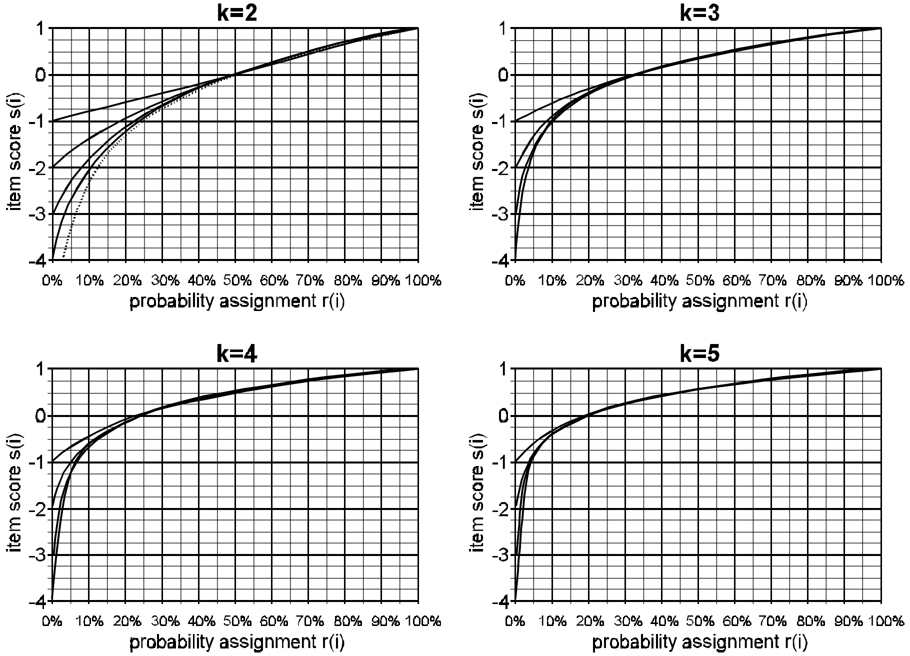


*Figure 8. Item scores as a function of probability assignment, for the four values of the tolerance parameter **t** given in Table 3.*
*The dotted line in the k=2 plot, below the T=4 curve, corresponds to the logarithmic rule with zero tolerance. For **k**>2 the zero-tolerance curves are indistinguishable from the T=4 curves for p>0.01.*

Note that the student must still attempt to maximize his score by selecting the best personal probability assignment; the only deviation involved is that this best possible assignment *r(i)* does not correspond exactly to the true probability *p(i)*.

It can be argued that the scoring rule will be better understood by students if we select the same value of **T** for all items, irrespective of the number of alternatives. For the purpose of this research we have used **T**=4, both for items where **k**=2 and where **k**=4; this corresponds to **t** = .0362 and **t** = .0010, respectively. In most practical applications, the deviation for **t** = 0.0362 (i.e. for **T**=4 and **k**=2) should prove acceptable.

4.2 Test score conversion

A further problem involves the final test score. The item scores range from some negative value (−4 in our case) to +1, through 0 points for unanswered questions (with a probability assignment 1/**k**). If the total test score is taken as the average of the item scores, or some multiple thereof, the result can lie between –40 and +10 for example. This is not common practice.

In our case, standard grades are on a scale of 1 to 10 with 1 being the lowest possible, and 10 the maximum. Furthermore, 6 .. 10 are 'pass' marks and 1 .. 5 mean 'fail', so the cutting score is 5.5; obviously, this cutting score should correspond to some ME score greater than 0 points since this can be achieved by leaving all questions unanswered.

Converting the ME test scores to our standard grades therefore requires a non-linear conversion rule, with the provision that a higher ME score must result in a higher final grade.

For the purpose of this research, we have applied the following exponential rule.

Defining: S1, S0 and Sc = maximum, minimum and cutting final grades,
M1, Mc and *ME* = maximum, cutting and actual ME test scores,

the final (converted) grade *S* is calculated as:

$$S(ME) = \frac{2\times(\mathrm{S1}-\mathrm{S0})}{1 + \left(2 \times \frac{\mathrm{S1}-\mathrm{S0}}{\mathrm{Sc}-\mathrm{S0}} - 1\right)^{\frac{\mathrm{M1}\,-\,ME}{\mathrm{M1}\,-\,\mathrm{Mc}}}} + \mathrm{S0} \qquad (4.4)$$

This rule defines a smooth conversion curve from ME scores to final grades, over the desired range from minimum to maximum final grades (S0 .. S1). The ME cutting score (Mc) can be set to correspond to any desired final cutting grade (Sc). Once these parameters are chosen the formula reduces to a simple form, giving the final test grade *S* as a function of the total ME score *ME*. In our case, with S1 = 10, S0 = 1, Sc = 5.5, M1 = 10 and Mc = 2.5, this results in:

$$S(ME) = \frac{18}{1 + 3^{\frac{10\,-\,ME}{7.5}}} + 1 \qquad (4.5)$$

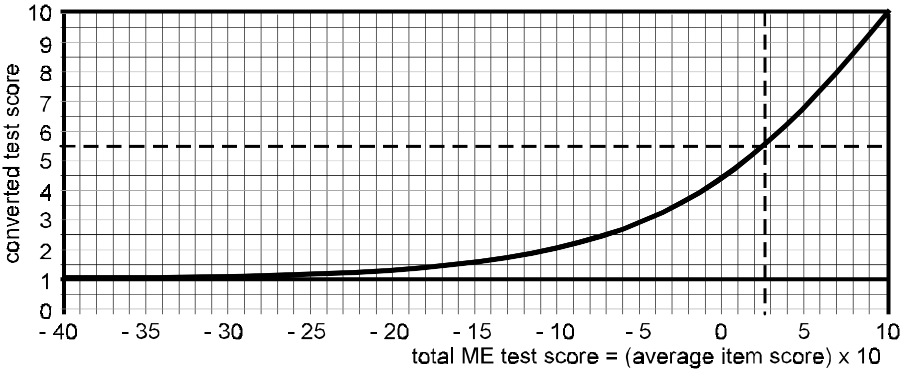


Figure 9. *Test score conversion: final (converted) test grade as a function of the ME test score.*

This conversion rule, illustrated graphically in Figure 9, proved acceptable in practice. The cutting score (Mc = 2.5) is set so that for **k** = 2, for example, a probability assignment of 60% to *all* correct alternatives (for all test items) will just lead to a pass. As can be derived from Figure 8, the 'cutting percentages' for **k** = 3, 4 and 5 are approximately 43%, 33% and 30% respectively. Note that these values are somewhat theoretical, merely giving an indication of the average 'knowledge' required. In the highly unlikely event that a candidate actually selects these percentages for the correct answers over the complete test, this would be considered an indication of extreme underestimation of his or her capabilities. This would lead to a major 'correction for lack of realism', as will be described later (section 4.4).

If we assume that all items are given the same probability assignment r by a perfectly realistic student, a proportion $p = r$ of these items will be right and $(1-p)$ will be wrong. In this case the 'cutting percentage' for **k** = 2 can be calculated on the basis of the expected total score $E(s)$ (i.e. the average item score times 10):

$$E(s) = [p \times s(p) + (1-p) \times s(1-p)] \times 10 = 2.5$$

This corresponds to $p = r \approx 78\%$. In other words, the cutting score for a **k** = 2 test is reached if the response for the correct alternative $r_c = 78\%$ for 78% of the items and $r_c = 22\%$ for the remaining items.

For **k** = 3, 4 and 5 the 'cutting percentages' are approximately 70%, 65% and 61% respectively, if we assume that the remaining probability is distributed equally over the remaining alternatives (for **k** = 3, for example: 70-15-15).

However, a probability assignment of 1/**k** for all alternatives (ME score = 0) will be converted to 4.38, which is rather high. For **k**=4, random guesswork (over a sufficiently large number of items) with an unweighted MC scoring rule would result in a total test score of (10–1)/**k** + 1 = 3.25 (on the same scale from 1 to 10). A modified (more 'severe') score conversion rule will be considered later (see section 6.2).

4.3 Measuring the degree of realism

Considering one item in a test, we can represent the student's uncertainty as to the correct answer by a probability distribution $\bar{p} = (p_1, p_2, ..., p_k)$ over the set {1, 2, .. , **k**} of possible answers, where $0 \le p_j$ and $\Sigma p_j = 1$. The student's response can be represented by $\bar{r} = (r_1, r_2, ..., r_k)$. For a perfectly realistic student the response $\bar{r}$ is equal to $\bar{p}$. In effect, such a student is stating: "Given my knowledge of the subject, this item is one of the many items for which my personal probability distribution is $\bar{p} = (p_1, p_2, ..., p_k)$. For these items, answer 1 is correct in p_1 cases, answer 2 is correct in p_2 cases, etcetera." To put this a different way: the response of a realistic student should correspond to his *personal* probabilities: for each alternative answer, he should indicate the likelihood of that particular alternative being correct - to the best of his knowledge. It is precisely this information that we wish to evaluate; therefore, we want our students to give a *realistic* response $\bar{r} = \bar{p}$.

The item score $s(i)$ is based on the response $r_c(i)$, the probability assigned to the *correct* alternative, and this score is maximized when $r_c(i) \approx p(i)$. By this means we attempt to ensure that the response is based on a *realistic* estimate of the personal probability that this is indeed the correct answer: it is in the best interest of the student to be realistic.

This is no easy requirement, however. The only straightforward probability assignments are those for ‘certainly true’ (100%), ‘certainly false’ (0%) and ‘certainly don’t know’ (1/**k**). We may be perfectly certain that two plus two equals four, and equally certain that two plus two does not equal five; we are also perfectly realistic in stating that we do not know whether a flipped coin will land heads or tails. But if we consider it ‘quite likely’ that Mozart lived in the eighteenth century, should we assign that statement a personal probability of 75%? Or 95%? When taking an ME test a student must translate his ‘feeling of uncertainty’ into a fairly accurate numerical response percentage, and this will obviously require some practice. (In some cases knowledge of statistics and probability theory may also be useful, as in “Three coins are tossed. The results are (a) three of a kind, (b) one pair, (c) other.”)

This problem is aggravated by the fact that most students are accustomed to Multiple Choice testing. Even in MC tests where some form of weighting is applied (assigning penalty points to incorrect answers), the best test-taking strategy is often to pick the most likely alternative - no matter how unlikely it may seem. Given the large number of questions in most MC tests, students may even get into the habit of skimming over ‘difficult’ questions and picking one of the alternatives almost at random - witness, in our experience, the clear tendency towards an approximately equal distribution of the responses over all alternatives when an item requires a longer than average calculation, regardless of the actual difficulty.
When faced with Multiple Evaluation testing, there is a marked initial tendency towards selecting the ‘most likely’ alternative, and assigning it an unrealistically high probability.

Initially, while students are still learning to respond with realistic probability assignments, some form of feedback is essential. Earlier research (DIRKZWAGER, [21]) has shown that even quite young children soon learn to be realistic when on-line feedback is provided. In order to provide this feedback, and for further analysis, some calculated measure of realism is required. DIRKZWAGER [22] discusses three possibilities: a least squares estimate, a z-score estimate, and a chi^2 estimate for calibration. However, the chi^2 distribution would not be sufficiently approximated in our case, since each test presented a total of only twenty options (as will be discussed later). For our research, the following least squares estimate of realism was considered suitable.

When a probability r is assigned to a large number R of alternatives (for many items) of which R_t are correct, we expect that the ratio $\rho = R_t/R$ equals r if the subject is realistic: $r(i) = p(i)$ for all i.
Assuming a linear relation between the true probabilities (p) and the probability assignments (r), $p = A\times r + B$, then summation over all $\mathbf{k}$ alternatives per item gives $B = (1–A)/\mathbf{k}$.
For a least squares estimate of A we therefore investigate the function:

$$\begin{aligned} f(A,r) &= \Sigma\Sigma\, R(r) \times (R_t(r)/R(r) - (A\times r + (1-A)/\mathbf{k}))^2 \\ &= \Sigma\Sigma\, R(r) \times ((R_t(r)/R(r) - 1/\mathbf{k}) - A\times(r - 1/\mathbf{k}))^2 \end{aligned}$$

(summation over all values of r, for all items)

Partial differentiation gives:

$$\delta f(A)/\delta A = 2A\times\Sigma\, R(r)\times(r-1/\mathbf{k})^2 \;-\; 2\times\Sigma\, R(r)\times((R_t(r)/R(r)-1/\mathbf{k})\times(r-1/\mathbf{k})$$

Setting $\delta f(A)/\delta A = 0$ and solving for A results in:

$$\mathbf{A} = \frac{\sum\sum \mathbf{R_t(\mathit{r})\times\mathit{r} - R_t(\mathit{r})/k - R(\mathit{r})\times\mathit{r}/k + R(\mathit{r})/k^2}}{\sum\sum \mathbf{R(\mathit{r})\times\mathit{r}^2 - 2R(\mathit{r})\times\mathit{r}/k + R(\mathit{r})/k^2}}$$

Since summation is over *all* observed values of r assigned to $\mathbf{m}\times\mathbf{k}$ options, i.e. over all $\mathbf{m}$ items in the test, we can simplify and rewrite this result on the basis of:

$\Sigma R(r) = \mathbf{m.k}$

$\Sigma R(r)\times r = \mathbf{m}$ (since $\Sigma r = 1$ for each item)

$\Sigma R_t(r) = \mathbf{m}$ (since one option for each item is correct)

$\Sigma R_t(r)\times r = \Sigma r_c$ (sum of probabilities assigned to the correct options)

$\Sigma R(r)\times r^2 = \Sigma\Sigma r^2$ (sum of probabilities for all options, for all items)

Therefore:

$$\mathrm{A} = \frac{\sum r_c - \mathbf{m/k} - \mathbf{m/k} + \mathbf{m.k/k^2}}{\sum\sum r^2 - 2\mathbf{m/k} + \mathbf{m.k/k^2}}$$

$$= \frac{\mathbf{k} \times \sum_{\mathrm{i=1}}^{\mathrm{m}} r_c(i) - \mathbf{m}}{\mathbf{k} \times \sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{k}} r(i,j)^2 - \mathbf{m}} \qquad (4.6)$$

With this estimate, we get the best least squares fit when we estimate the true probabilities with the linear formula

$$p = \mathrm{A}\times r + (1\text{-A})/\mathbf{k}. \qquad (4.7)$$

If A = 1, the subject is perfectly realistic; A < 1 corresponds to overconfidence (assigning too extreme probabilities) and A > 1 corresponds to underconfidence.
In our research, feedback to the students for each test was based on this least squares method.

4.4 Final score corrected for lack of realism

For diagnostic testing, the fact that a student must attempt to be realistic about his (lack of) knowledge is a valuable aid to learning. For most certifying tests however, where the test result has a major impact on the student's progress, realism should not be such a determining factor. This implies that some correction for lack of realism is required. On the other hand, full compensation for lack of realism is undesirable (even if it were possible) since that would remove all incentive to assigning realistic probabilities. As in most practical applications, a compromise is required: sufficient compensation to ensure that the test result is determined to a large extent by the student's knowledge of the subject being examined, while maintaining 'realism' as a sufficiently important factor to discourage over- and underestimating.

This correction for lack of realism can be achieved in two ways. We can modify the tolerance parameter **t** for each individual student (“t-correction”), or we can modify the probability assignments *r* (“r-correction”). These two methods are described below.

t-correction:

Since the tolerance parameter **t** already provides a degree of ‘protection’ for overconfident students, we can consider modifying this parameter to further compensate for lack of realism. Using the least squares estimate for realism (A), we have seen that we may estimate the true probabilities with the linear formula $p = A\times r + (1-A)/\mathbf{k}$ (4.7).

This equates to a corrected probability estimate $p = (1 - \mathbf{t'}\times\mathbf{k})\times r + \mathbf{t'}$, if we let the tolerance parameter equal $\mathbf{t'} = (1-A)/\mathbf{k}$. Provided a student is completely consistent (so that A is constant for all items), substituting this value for the tolerance parameter in the scoring rule will result in item scores that reflect ‘perfect realism’. This is true, in theory, both for overconfident and for underconfident students - *provided they are completely consistent* in their lack of realism.

However, this correction can actually penalise realism or underestimation by setting a smaller (more severe) or even negative value for the tolerance parameter. If a student is not consistently underestimating, a low probability assignment for one or more items will therefore lead to even more extremely negative item scores - possibly even minus infinity. For this reason we must set a lower limit for this parameter, and a simple choice is the original value for **t**. Unfortunately, this implies that there will be no correction for underestimation.

On the other hand, this t-correction may be too all-forgiving for overconfidence. The maximum penalty (for a probability assignment of 0% for the correct alternative) could become negligible, thereby degrading the scoring rule to a complicated variation on Multiple Choice. For this reason we must also limit the maximum value of the corrected tolerance parameter, to maintain a maximum penalty of at least -1 for instance.

On the basis of the values given in Table 3, the above considerations result in the tolerance parameter ranges $0.0362 \le \mathbf{t'} < 0.5$ for $\mathbf{k} = 2$, and $0.001 \le \mathbf{t'} \le 0.0833$ for $\mathbf{k} = 4$. Even then, this correction is possibly too lenient: it has no effect for 'correct' assignments (r > 1/**k**) while reducing the penalty for 'incorrect' assignments (see Figure 10).

r-correction:

An alternative method is to 'correct' the reported probability assignments according to the same formula, $r' = p = A\times r + (1-A)/\mathbf{k}$. For values of $A < 1$ (i.e. overconfidence, on average over the entire test) the effect is to shift *all* probability assignments to some value closer to 1/**k**; for $A > 1$ (underestimating on average) the effect is to shift *all* probability assignments outwards towards 0% or 100% - or even to negative values or values greater than 100%. In the latter cases, the corrected values for the probability assignments must be restricted to the range $0\% \leq r' \leq 100\%$.

As before, we do not wish to overcompensate and so A is restricted to the range $0.5 \leq A \leq 2$ for **k** = 2 or $0.75 \leq A \leq 2$ for **k** = 4 before performing this correction. These ranges were chosen to compensate for what we considered acceptably realistic responses. For **k** = 2 an overconfident student could give a response r = 0% when the true probability was actually p = 25%, and r = 100% for p = 75%; conversely, an underconfident student's responses could be r = 25% for p = 0%, and r = 75% for p = 100%. For **k** = 4 underconfidence is treated in a similar fashion (compensation up to r = 12.5% for p = 0% and r = 62.5% for p = 100%), but overconfidence is more readily penalised: r = 100% for $p \approx$ 80% is borderline-acceptable and this corresponds to r = 0% for $p \approx$ 6%. The difference between the two settings (more tolerance for overconfidence for **k** = 2 than for **k** = 4) was based on practical experience during the first few tests: students proved consistently more realistic when considering *all* alternatives carefully for the **k** = 4 items.

The second correction method (adjusting the probability assignment) seems preferable, since it corrects for both over- and underconfidence as illustrated in Figure 10. In our initial research, however, we have taken the higher of the two corrected scores as the final test score for the students.

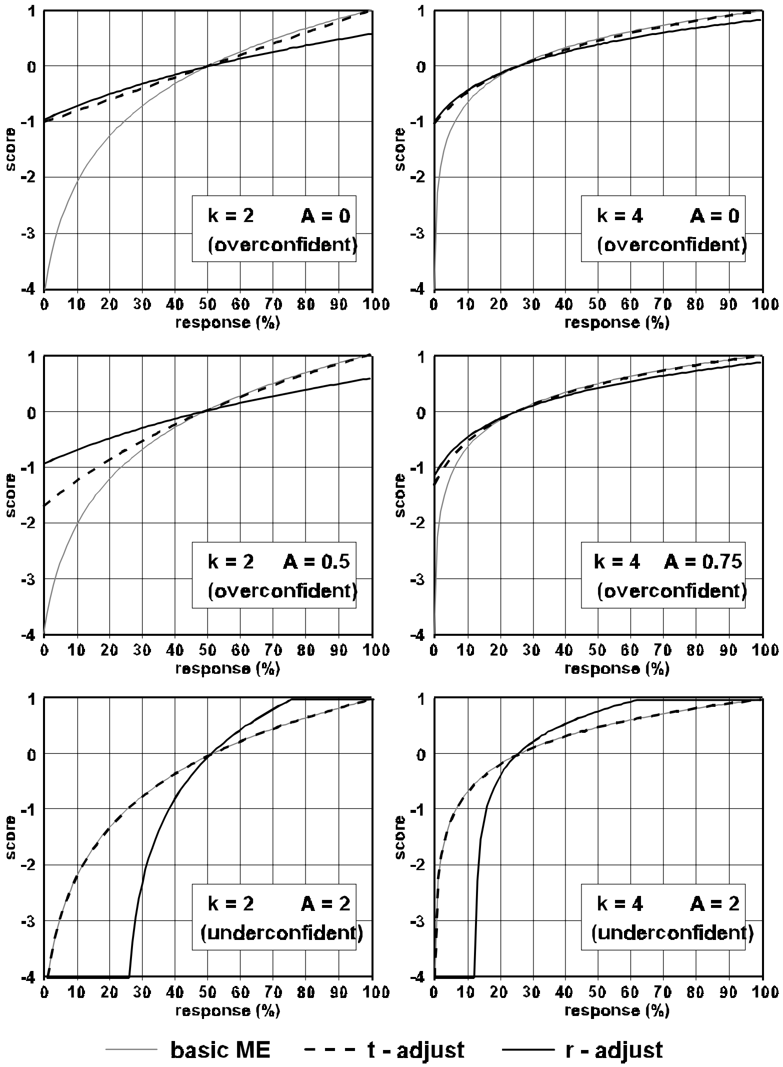


Figure 10. *Correction of item scores for lack of realism. The basic ME item score runs from -4 to +1. Adjusting the tolerance parameter t has negligible effect for responses above 1/**k** and only corrects for overestimating. Adjusting the response r also corrects for underestimating, and reduces the score for responses above 1/**k** for overconfident students.*

It should be noted that, certainly when using the second method, some reward for realism is maintained in all cases – not only due to the range restrictions applied. The realism score (A) is based on *all* probability assignments, for all items in the total test. If a student is overestimating his probabilities on average, *all* his probability assignments will be reduced - including those where he was indeed 100% certain of the correct answer, thus reducing the score for those items. Conversely, all responses of a student who is underestimating on average will be taken as more extreme - including those with a response $r < 1/\mathbf{k}$ for the correct alternative, which will be taken as closer to (or even equal to) 0%, thus increasing the penalty. In both situations, the final (corrected) test result will be lower than that which this student would have obtained if he were perfectly realistic.

4.5 Derivation of 'MC scores' for comparison

One of the main objectives of our research was to evaluate Multiple Evaluation and determine whether it could present a more reliable testing tool than Multiple Choice. To this end we must also obtain MC scores for the same group of students, the same subject matter and preferably the same or exactly comparable test items.
Since this research was carried out in a regular educational environment and was not allowed to disrupt regular activities, a practical approach was to derive estimated MC choices from the ME probability assignments.

At first sight this would appear a simple matter. If the probability assignment for the correct alternative is higher than the assignments for all incorrect alternatives, the item can be scored as 'correct' (item score $s(i)$ =1) and otherwise it is 'incorrect' ($s(i)$ = 0). In fact exactly this rule was applied to derive one set of 'hard' or 'maximum probability' Multiple Choice test scores, referred to as MC1.
On more careful consideration one may doubt the validity of this simple rule, however. When a student assigns 100% (or 0%) to the correct alternative, an item score of 1 (or 0) would appear perfectly appropriate. But what if, given four alternatives, the student assigns 50% to *two* alternatives of which one is correct (and 0% to the other two incorrect options)? Or, to take a slightly more extreme

case: 40% to the correct alternative and 60% to an incorrect one? In a Multiple Choice situation, which of these two alternatives would the student have picked? These considerations led us to an alternative, more gradual rule which results in a second set of 'uncertainty-scale' scores, referred to as MC2. The ME probability assignments are scored in three distinct groups: definitely correct, definitely incorrect, and dubious:

if $r(i) > \mathrm{x}$ then $s(i) = 1$, where $\mathrm{x}_{k=2} = 0.75$ and $\mathrm{x}_{k=4} = 0.6$;
if $r(i) < \mathrm{y}$ then $s(i) = 0$, where $\mathrm{y}_{k=2} = 0.25$ and $\mathrm{y}_{k=4} = 0.13$;
if $\mathrm{y} < r(i) < \mathrm{x}$ then $s(i) = r(i)$.

By this rule, a high probability assignment to the correct alternative (more than halfway up the scale between 1/**k** and 100%) is taken as a correct answer and a very low probability assignment is taken as incorrect. In other words, when a student has apparently made a clear choice for or against the correct alternative the item score is 1 or 0 - exactly as in the 'hard' MC1 version. In the dubious range (to both sides of 1/**k**), however, we use the linear probability scoring rule: the probability assignment *r(i)* is taken as the item score *s(i)*. Taken over a sufficiently large number of questions, this mid-range score could therefore be considered an 'expected MC' score.

One can argue that even this modified rule is questionable, and that some form of weighted random item scoring should be applied. Given the small number of items in our tests, however, random scoring would lead to wildly erratic results. The two rules described above both have the advantage that they can be consistently compared to the actual ME scores.

It should be noted that the 'estimated uncertainty-scale' (MC2) is not intended as a true scoring rule for determining valid test scores. As such it would be completely unsuitable, as SHUFORD [73] proved: the optimal test-taking strategy for students would be to pick the most likely alternative (no matter how unlikely that may seem) and assign a probability of one to that option (and zero to all others). In other words, intelligent students would soon degrade such a rule to traditional pick-one Multiple Choice. Its only application here is to estimate a hypothetical MC score on the basis of the probabilities assigned during a Multiple Evaluation test.

4.6 Test reliability

On the basis of earlier research (Chapter 2) we concluded that the test reliability of Multiple Evaluation is at least equal to if not far better than that for traditional Pick One. Initially, therefore, our prime interest was the practical feasibility of ME testing. Test validity came in a close second, since earlier researchers often expressed doubts concerning this aspect. No problems were expected concerning test reliability as such.

However, it was our intention to use extremely short tests. The length of our traditional 4-option MC tests is based on 10 items per 50-minute period, and for the purposes of this research we wished to evaluate both computer-based testing (4-option) and paper&pencil (2-option). Since these tests must both be completed within a single 50-minute session, we would be restricted to 5-item 4-option and 10-item 2-option tests. The reliability of such short tests seemed doubtful, but the semester-end grade was to be based on the overall average (after 10 or more tests) and even in the worst case this should prove reliable. Furthermore, we intended to derive estimated MC scores (as described earlier) and compare the reliability of ME testing versus MC.

For these reasons, an acceptable numerical measure for the reliability of ME tests is required as a first estimate of the reliability of the weekly results. Since we wish to compare the results to those obtained on the basis of (estimated) MC scores, this numerical measure should preferably be one that is comparable to the well-known Cronbach’s alpha. This proved a greater hurdle than expected.

In essence, test reliability can be defined as test-retest reliability. Suppose that we could present each student with a test and calculate his final score, then brainwash him of all knowledge concerning this test and retest him with exactly the same items and calculate a second final score. The correlation between the two scores would then give a good indication of the test reliability: in the ideal case, with perfect reliability, the two scores should be identical. This leads to the definition of test reliability as ρ_{tt} : the correlation between test and retest.

In a practical environment we cannot brainwash the students. The two tests must differ, therefore, but they should both test the same abilities to the same degree. In our situation, parallel testing was not possible. Quite apart from the difficulty of constructing two such similar tests, we were restricted to one 50-minute lecture period per week for all tests.

A well-known variation on this approach is the split-half estimate of reliability: a test is split into two equal halves (e.g. odd and even items) and the correlation between the two resultant scores is computed. However, HOYT (1941, [47]) gives sound reasons against split-half estimates, and for our extremely short tests (only five or ten items) this method must also be rejected.

Going back to basics, Hoyt continues “The coefficient of reliability of a test gives the percentage of the obtained variance in the distribution of test scores that may be regarded as true variance, that is, as variance not due to the unreliability of the measuring instrument.” Using variance decomposition, he derives an estimate for the coefficient of reliability. It can be shown (Sirotnik, 1970) that Hoyt’s result equates to the well-known ‘coefficient alpha’ measure for test reliability proposed by Cronbach (1951):

$$\alpha = \frac{\mathbf{m}}{\mathbf{m} - 1} \times \left(1 - \frac{\sum_{i=1}^{m} \sigma^2_{X_i}}{\sigma^2_X} \right) \tag{4.8}$$

where **m** is the number of items, $\sigma^2_{X_i}$ is the item score variance of item *i*, and σ^2_X is the test score variance.

When only item scores of 1 or 0 are possible, as for MC, this formula reduces to Kuder and Richardson’s KR-20. This approach should therefore lead to a good basic comparison between traditional MC scoring and Multiple Evaluation scoring, provided a similar estimate can be derived for our Multiple Evaluation tests. This requires some theoretical analysis.

Let us first reconsider the approach outlined above. The test reliability estimate is based on the sum of the variances of the item scores ($\Sigma\sigma_{Xi}^2$) and the variance of the overall test scores (σ_X^2). However, we may assume that the component associated with the item scores in a test will also depend to some extent on each individual student. This implies that the item score variance 'between students', σ_{Xi}^2, includes some ability-dependent interaction variance 'between students and items'. We consider this an acceptable assumption, given inhomogeneous items in a real test. The traditional calculation of item variance however, when calculating Cronbach's alpha, includes this ability-dependent variance as part of the variance in the responses. This leads to a lower estimate of the test reliability, but it is unavoidable when no better estimate is available, as is the case in traditional MC tests. For Multiple Evaluation tests another estimate will be derived in the following section; in our empirical study, however, we will also use the more traditional calculation in order to obtain a 'fair' comparison with MC tests.

Theoretically, test reliability is based on the correlation between true scores and observed scores: $\rho = \sigma_T^2/\sigma_X^2$. Since we do not know the true scores, obviously, σ_T^2 must be estimated. This estimate can be based on the test score variance σ_X^2 and the error variance σ_E^2: $\sigma_T^2 = \sigma_X^2 - \sigma_E^2$. Test reliability is then estimated as:

$$\rho = (\sigma_X^2 - \sigma_E^2) / \sigma_X^2 = 1 - (\sigma_E^2 / \sigma_X^2) \qquad (4.9)$$

We must now consider how these variances can be calculated, or at least estimated, from actual test results.

The total test variance (σ_X^2) is estimated by the variance of the observed final scores. The error variance (σ_E^2) requires more careful consideration when using probability measurement in general, and Multiple Evaluation in particular. In the next section we will therefore explore the theoretical reliability of Multiple Evaluation tests on the basis of a model, and evaluate the results of this analysis by means of computer simulation. This approach allows us to calculate theoretical test-retest reliabilities, since a computer simulation can generate any desired number of items and (modelled) responses.

One further complicating factor for the ME scores is the students' (lack of) realism. For this reason our final test scores are based on a correction for realism, as described in section 4.4. We first determine 'realistic' probability assignments *p(i,j)*, as far as possible, based on the realism score (A) and the formula $p = A \times r + (1-A)/\mathbf{k}$. When estimating test reliability we will use these corrected probability assignments. The dominant effect of this correction is on the most negative item scores: for overestimators the penalty scores are reduced and for underestimators they will increase. This effect can be considerable, due to the logarithmic scoring rule. Since earlier research predicts that students tend towards overestimation, we may expect that this correction will reduce the item-score variances considerably. If the effect on the (variance of the) final scores is less, which we tend to expect, the resulting estimate of test reliability will improve. We consider this acceptable, since our actual final grades are also calculated on the basis of these corrected probability assignments.

4.7 Test reliability: modelling and simulation

As is the case with any other test, the reliability of probability measurement scores can be defined as the correlation between the scores obtained in two independent administrations of the same test.

Let us consider a student taking such a test. For each item, his basic answer is a subjective probability distribution of the response alternatives and the score is a function of the subjective probability assigned to the correct alternative. If students were willing and able to express their subjective probability with absolute accuracy, they would do this at each occasion. As a result, the reported subjective probabilities and hence the obtained scores would be the same at each test administration. This would yield a reliability coefficient of one, which of course nobody would believe. Assuming that the test takers are willing to convey their subjective probability, one must accept that they are not able to do so with absolute accuracy, but that the reported vector $\mathbf{r} = (r_1, .., r_k)$ deviates from their true subjective probability vector $\mathbf{p} = (p_1, .. , p_k)$. This means that the perception of the subjective probability is blurred in some way, causing variability of the reported vectors **r** across test administrations, and it is precisely this variability which causes an imperfect reliability.

To formalize these ideas, the approach proposed by Verstralen and Verhelst (2000, [87]) is followed, with some modifications. Let us first consider the variability of a reported subjective probability for one subject and one item. The basic assumption is that at any presentation of an item, the reported vector **r** is a random draw from a probability distribution with expected value **p**. A suitable distribution is the Dirichlet distribution (a multivariate generalization of the beta distribution). In our case, since the score only depends on the probability assigned to the correct alternative r_c, we can suffice by considering the marginal probability density function of r_c, the beta density function:

$$f(r_c) = \frac{\Gamma(\mu)}{\Gamma(\mu p_c) \times \Gamma[\mu(1-p_c)]} \times r_c^{\mu p_c - 1} \times (1-r_c)^{\mu(1-p_c)-1} \qquad (4.10)$$

where μ (> 0) is an 'accuracy' parameter. For a beta distributed random variable, expected value and variance are well known. In the present case:

$$E(r_c) = p_c \quad \text{and} \quad Var(r_c) = \frac{p_c(1-p_c)}{\mu + 1}$$

There is an inverse relationship between the variance of r_c and the parameter μ: if μ tends to infinity, the variance tends to zero, and if μ tends to zero the variance of r_c tends to the variance of a binomial distribution $B(\mathbf{m}, p_c)$ with $\mathbf{m} = 1$.

To investigate the theoretical reliability of a test score, we must next account for the variability of the true subjective probability across items and across students. Let us assume that the true subjective probability depends on an underlying (latent) variable θ which may differ from person to person, and on a difficulty parameter β which may vary from item to item. In other words, we wish to construct an item response model that expresses the relation between the latent variable θ and the true subjective probability p_{ic}. A quite general model, proposed by Verstralen and Verhelst, is given by:

$$p_{ic}(\theta) = \frac{\gamma_{ic}[1 + \exp(\theta - \beta_i)]}{\gamma_{ic}[1 + \exp(\theta - \beta_i)] + \sum_{j \neq c}^{k} \gamma_{ij}}$$

where β_i represents the difficulty parameter of item i and γ_{ij} is an attraction parameter for the j-th alternative.

The latent variable θ is unbounded. Total ignorance corresponds to a θ-value approaching minus infinity, and in that case:

$$\lim_{\theta \to -\infty} p_{ic}(\theta) = \frac{\gamma_{ic}}{\sum_{j=1}^{k} \gamma_{ij}}$$

This should correspond to a uniform distribution of the subjective probabilities, $p(-\infty) = 1/\mathbf{k}$ for all alternatives. This is achieved by setting $\gamma_{ij} = 1/\mathbf{k}$ for all alternatives and all items, which results in a simplified IRT-model:

$$p_{ic}(\theta) = \frac{1 + \exp(\theta - \beta_i)}{\mathbf{k} + \exp(\theta - \beta_i)} \qquad (4.11)$$

The final step is to assume a distribution function of θ to account for the variability across persons. A simple choice is to assume the normal distribution.

For binary responses we can easily construct an IRT model which is in full correspondence with this continuous response model. The probability of a correct response is simply set equal to the subjective probability of the correct alternative. Representing the binary response on item i by X_i, this gives

$$p(X_i = 1 \mid \theta) = p_{ic}(\theta) \qquad (4.12)$$

The score on a test with binary responses is just the number of correct responses.
It should be noted that the transition from the continuous model to the binary model is consistent with the assumption of realism: the expected number of correct

binary answers equals the sum of the subjective probabilities assigned to the correct alternatives.

Simulation

The relation between the reliability of probability measurement and binary scoring was investigated by means of a small simulation study, using the IRT models derived above.

The number of alternatives per item is set at **k** = 4 and the number of items in the test is either 10 or 20. For this simple study all item parameters β_i are set equal to zero. For the continuous model the logarithmic scoring rule (3.3) is used with the tolerance parameter **t** = 0.0010, corresponding to T = 4 (Table 3, p. 44). The accuracy parameter μ is given the values 3(1)10, 15, 20, 30 and 40, giving twelve estimates of the reliability in each cell. A θ-value is drawn at random from the standard normal distribution for each simulated person, and the subjective probability p_{ic} is computed for each item according to our model (4.11). Based on this subjective probability two independent values r_{ic} are drawn from the beta distribution (4.10) for each item and two corresponding item scores are calculated (3.3); these scores are summed over items to produce two independent test scores, thus simulating a test-retest situation. To get stable estimates of the reliability, this procedure was repeated for **n** = 20000 simulated persons. The reliability estimate is the correlation between the two series of **n** test scores.

For the binary scores two minor modifications to the model (4.12) are required. Instead of using the true subjective probability p_{ic} as the probability of a correct response, the sampled values r_{ic} are used. A uniformly distributed random number v in the (0, 1) interval is drawn, and the binary response is considered correct if v is less than or equal to r_{ic}. As in the continuous case, the two series of test scores are correlated, giving an estimate of the reliability of the binary scoring.

The results of this simulation study are given in Figure 10.

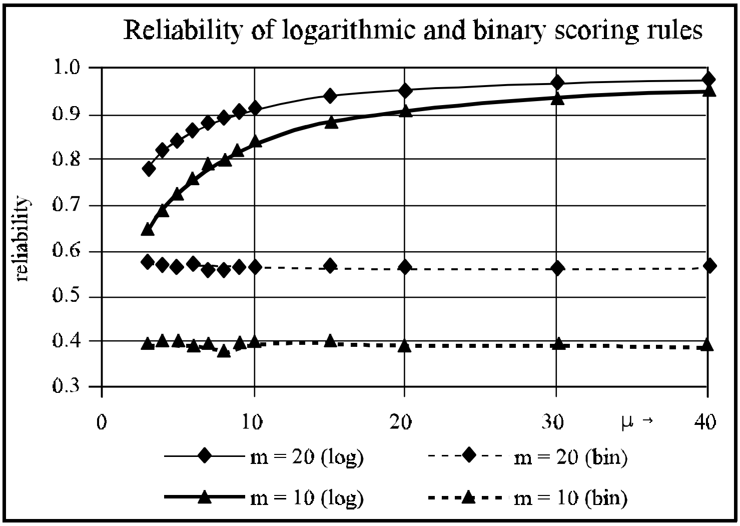


Figure 10. Test-retest reliability estimates as a function of the accuracy parameter μ, on the basis of computer simulation.

Several features are noteworthy:

- The relation between the accuracy parameter μ and the reliability is important. As predicted, when μ approaches zero the reliability approaches the reliability of the binary scoring model. The initial increase in reliability as a function of μ is quite rapid, then slows down. When μ approaches infinity, the reliability approaches one.
- For the binary scoring model, using r_{ic} as the probability of a correct response, the value of μ does not have any noticeable influence.
- For the binary model, an estimate based on the true subjective probabilities p_{ic} yields 0.39 for the 10 item test and 0.56 for the 20 item test. This is in accordance with the Spearman-Brown formula: 2×0.39/(1+0.39) = 0.56. The same Spearman-Brown relation holds for the continuously scored tests at each value of μ.
- The difference between continuous scoring and binary scoring is substantial, even for $\mu = 3$: 0.64 versus 0.39 for the 10 item test, and 0.78 versus 0.56 for the 20 item test.

Although the results of this simulation are promising, we must stress that actual values of the parameters will have a significant influence on empirical results. The accuracy parameter μ is very important, as illustrated above. The same applies for the latent variable θ; both its variance and its mean will influence the reliability. The higher θ, the skewer the distribution of the subjective probabilities will be, and hence the smaller the variance of the observed (reported) probabilities. This relationship is displayed in Figure 11, for two values of the standard deviation sd(θ) and two values of μ.

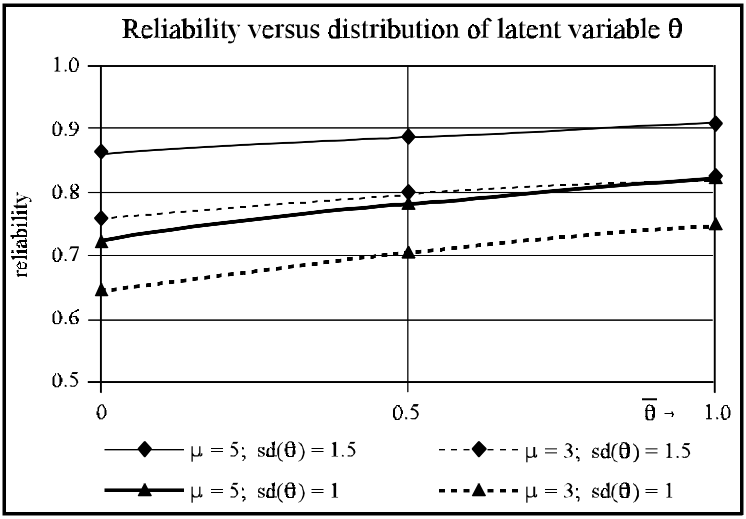


Figure 11. Test-retest reliability estimates as a function of the latent 'ability' variable θ, for two fairly low values of the accuracy parameter μ.

So far we have explored test-retest reliability for basically identical tests. In order te draw more general conclusions we can analyse a hypothetical experiment.

Suppose the true subjective probability vector of a student for a **k** = 4 item equals $\mathbf{p}_i$ = (0.6, 0.3, 0.1, 0.0). Now imagine the (in principle infinite) set of items which have exactly the same subjective probability distribution for this student. Since $\mathbf{p}_i$ is his *true* subjective probability, alternative A must be correct for 60% of these items, alternative B for 30%, alternative C in only 10% of the cases and alternative

D is never correct. From the viewpoint of the student this amounts to the situation that he has to consider the specific item he is responding to as being randomly drawn from this set. This experiment is consistent with the notion of realism.

We assume that this situation holds for all students and all items, with differing values for $\mathbf{p}_\mathrm{i}$. Now suppose we administer two tests of equal length, constructed in such a way that for each student the $\mathbf{p}_1$-vector for the first items in both tests is the same, the $\mathbf{p}_2$-vector is the same for the second items, and so on. This means that each student has an identical set of **p**-vectors in both tests, $\{\mathbf{p}\}$. The two scores will not be identical, however: even though $\mathbf{p}_\mathrm{i}$ is identical for each pair of twin items, the correct alternative may differ.

The correlation between the two series of scores can be derived from the variance decomposition theorem. Let S be the score for either of the two tests, then:

$$\mathrm{Var}(S) = \mathrm{Var}[E(S\,|\{\mathbf{p}\})] + E[\mathrm{Var}(S\,|\{\mathbf{p}\})]$$

$\mathrm{Var}[E(S\,|\{\mathbf{p}\})]$ is the variance of the true scores and $E[\mathrm{Var}(S\,|\{\mathbf{p}\})]$ is the expected variance of the scores. The interpretation of the latter term is important: as described above, it refers to the variance of the scores across replications of the test with different items but the same **p**-vector for each item pair. The correlation between two series of test scores S and S' is then given by

$$\rho(S,S') = \frac{\mathrm{Covar}[E(S|\{\mathbf{p}\}),\ E(S'|\{\mathbf{p}\})]}{\sigma_S \times \sigma_{S'}} = \qquad (4.13)$$

$$= \frac{\mathrm{Var}[E(S|\{\mathbf{p}\})]}{\mathrm{Var}[E(S|\{\mathbf{p}\})] + E[\mathrm{Var}(S|\{\mathbf{p}\})]}$$

Indicating the test score as S, the score on item i as s_i and the score if alternative j is the correct one for item i (which occurs with probability p_{ij}) as s_{ij} , we find:

$$E(s_i\,|\,\mathbf{p}_\mathrm{i}) = \Sigma_j\, p_{ij} \times s_{ij} \text{ and}$$

$$E(S\,|\,\{\mathbf{p}\}) = \Sigma_i\, E(s_i\,|\,\mathbf{p}_\mathrm{i})\,; \qquad (4.14)$$

$$\mathrm{Var}(s_i\,|\,\mathbf{p}_\mathrm{i}) = \Sigma_j\, p_{ij} \times [s_{ij} - E(s_i\,|\,\mathbf{p}_\mathrm{i})]^2 \text{ and}$$

$$\mathrm{Var}(S\,|\,\{\mathbf{p}\}) = \Sigma_i\, \mathrm{Var}(s_i\,|\,\mathbf{p}_\mathrm{i}) \qquad (4.15)$$

Note that $\rho(S,S')$ expresses the correlation between test scores of randomly composed tests with identical **p**-vectors but different items. Therefore it is not the reliability of the test; it is a generalization coefficient with respect to a universe of items, where the sampling of items is done as described above. As such, it can be considered as a lower bound of the reliability: $\rho(S,S')$ will never be higher than the reliability.

If the true subjective probabilities $\mathbf{p}_i$ were known the formulae (4.14) and (4.15) could be evaluated directly. However, we must accept that the observed response vectors $\mathbf{r}_i$ will contain a measurement error. The actual score is a function of $\{\mathbf{r}\}$, and this makes it more difficult to derive a good estimate of $\rho(S,S')$. As a simple approximation we can replace the vector $\mathbf{p}_i$ in the formulae (4.14) and (4.15) by the observed vector $\mathbf{r}_i$ in a simulation study and examine the relationship between $\rho(S,S')$ and the reliability.

Figure 12 shows this relationship for two values of the accuracy parameter μ.

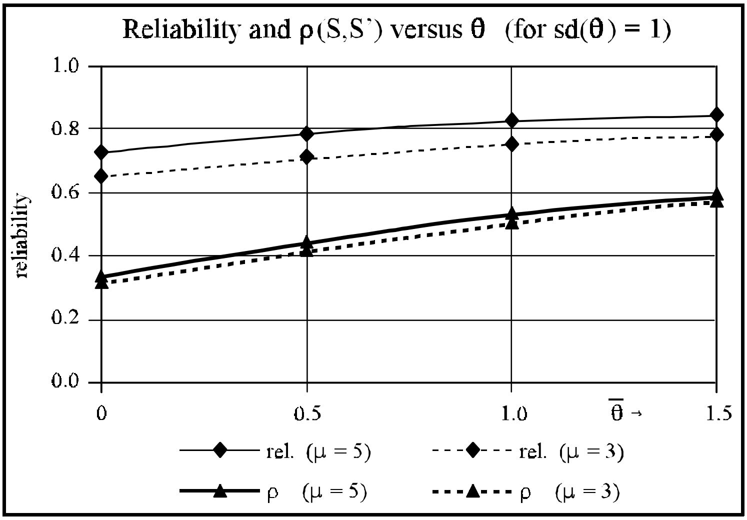


*Figure 12. Computer simulation allows us to compare 'true' test-retest reliability and a lower-bound estimate ρ using the observed response vectors **r**. The results clearly show that ρ is a quite conservative estimate of the true test reliability.*

The standard deviation of the latent variable θ is one and, as before, the results are estimated for a simulated sample of 20000 observations. It is immediately apparent that $\rho(S,S')$ provides an extremely conservative and cautious estimate of test reliability. Even at the highest θ-value simulated, a value of less than 0.6 corresponds to a test reliability of approximately 0.8.

4.8 Test reliability: numerical estimation

For our empirical results test-retest or split-half estimates are not feasible, so we must rely on an analysis based on item and test variance. One option is to use Cronbach's alpha:

$$\alpha = \frac{\mathbf{m}}{\mathbf{m}-1} \times \left(1 - \frac{\sum_{i=1}^{m} \sigma^2_{X_i}}{\sigma^2_X}\right)$$

This can be used to estimate the reliability of our three scores:

- ME: Multiple Evaluation with correction for realism (see section 4.4);
- MC1: 'hard' estimated Multiple Choice scores, 1 or 0 only;
- MC2: 'uncertainty-scale' estimated Multiple Choice scores.

No matter which of these scoring rules we consider, the number of items **m** remains the same and the overall test score variance (σ^2_X) is calculated from the corresponding observed final scores. The overall variance within items ($\Sigma\sigma^2_{Xi}$) is the sum of the individual item score variances.

For the MC1 scores, only item scores of 1 or 0 are possible. Therefore $p_c(i)$ is the average of s_i over all **n** students, with item variance $\sigma^2_{Xi} = p_c(i) \times (1 - p_c(i))$. For the MC2 scores, where $0 \leq s_i \leq 1$, σ^2_{Xi} is the item variance between all **n** students, for each item.

The same method can also be used for the ME scores, but in the previous section we have derived another estimate (4.13) for the lower bound of the reliability of Multiple Evaluation tests. Using the expected score $E(S|\{\mathbf{p}\})$ with its variance $\mathrm{Var}(S|\{\mathbf{p}\})$ as defined in formulae (4.14) and (4.15), but replacing $p(j)$ by the corrected-for-realism response *r(j)* in these calculations as discussed earlier, we can define an estimate of ME test reliability as:

$$\textbf{rho1} = \frac{\mathrm{Var}[E(S)]}{\mathrm{Var}[E(S)] + E[\mathrm{Var}(S)]} = 1 - \frac{E[\mathrm{Var}(S)]}{\mathrm{Var}[E(S)] + E[\mathrm{Var}(S)]} \qquad (4.16)$$

Alternatively, the denominator in (4.16) can be estimated by the observed score variance σ^2_X to obtain:

$$\textbf{rho2} = 1 - \frac{E[\mathrm{Var}(S)]}{\sigma^2_X} \qquad (4.17)$$

or:

$$\textbf{rho3} = \frac{\mathrm{Var}[E(S)]}{\sigma^2_X} \qquad (4.18)$$

The second estimate of ME test reliability, **rho2**, was derived independently by DIRKZWAGER (2001, [27]).

At this point it is unclear which of these options is to be preferred. For this reason some further simulations were used to compare the various estimates, for test scores after correction for realism. Smaller numbers of students were set, with wider spreads of the various parameters to create less reliable tests. Some results are shown in Figure 10. It is immediately apparent that the reliability estimates that depend on σ^2_X (i.e. $\rho2$, $\rho3$ and α) fluctuate over a wide range.

There is a simple explanation for this effect: correction for realism tends to move extremely incorrect responses closer to 1/**k**, giving scores close to zero, and tends to increase the higher scores for 'underestimators'. These two effects, combined, lead to higher values for σ^2_X so that rho2 increases and rho3 is reduced, as shown in Simulation 1. On the other hand, reducing the value of β and increasing μ leads to higher scores in general, so that all scores tend to increase; σ^2_X becomes smaller so that rho2 decreases and rho3 increases, as shown in Simulation 2.

With such a small number of students all reliability estimates are rather unstable, however; Simulation 3 is for 5000 students, using the same parameters as Simulation 2. This reduces the fluctuations, giving a clearer picture.

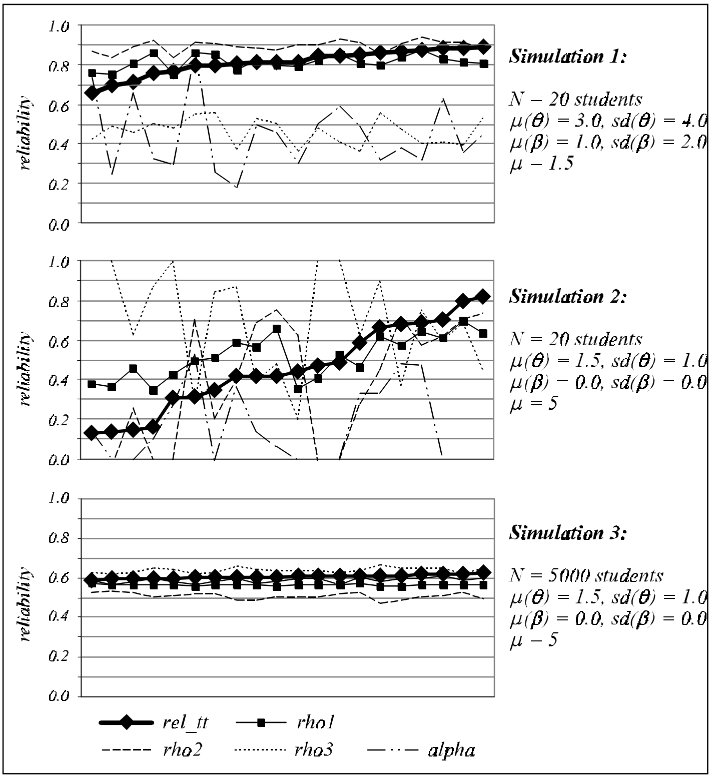


Figure 10. Reliability estimates for simulated tests. In general ρ1 gives the closest approximation to r_tt; α, ρ2 and ρ3 are influenced to a greater extent by the simulation parameters.

Based on these simulations, we provisionally consider that ρ1 is to be preferred. This is especially true when testing relatively small numbers of students, as the Simulations 2 and 3 illustrate: all parameters were the same in both cases, so that we may assume that the reliability of these tests should be almost identical. For the N=5000 run, r_tt ≈ 0.61 and ρ1 ≈ 0.57; in the N=20 run, ρ1 varies from 0.35 to 0.70 with an average of 0.55.

Our empirical results, to be discussed in the next chapter, were obtained for a group of approximately 60 students. Our main conclusions concerning test reliability will be based on ρ1, but we will also calculate α, ρ2 and ρ3.

On the basis of the estimated reliability (α_{MC} or $\rho 1_{ME}$) of an actual test, we can estimate the number of items (**M** = **L**×**m**) required for a desired reliability α_d by means of the Spearman-Brown formula:

$$\alpha_d = \frac{\mathbf{L} \times \alpha}{1 + (\mathbf{L} - 1)\alpha}$$

which can be re-written as:

$$\mathbf{L} = \frac{1 - \alpha}{\alpha} \times \frac{\alpha_d}{1 - \alpha_d} \qquad (4.19)$$

For a desired $\alpha_d = 0.75$, for example, **M**0.75 = **L**×**m** = [3×(1-α) / α] × **m**. Current practice is to aim at this value (α = 0.75) for end-of-course MC examinations, although it is not always achieved.

Finally, we can consider some other possible numerical estimates that may be used for test evaluation.

We can estimate of the difficulty of the items, in relation to the group of students involved. On average, we would hope that the students will assign the greatest probability to the correct answer and lower probabilities to incorrect alternatives. One indication of this is provided by the average (uncorrected) ME scores $s(i,j)_{av.}$ for all alternatives *j* for all items *i*. The best items should exhibit some (but not too extreme) consensus regarding the correct answer and similar slightly negative scores for all incorrect alternatives. If the average scores for incorrect options are too negative, the item may be considered too easy. Note that these average scores are mainly determined by students who feel quite confident about their selection: probability assignments near 1/**k** have relatively little effect.

Alternatively we can evaluate the actual probability assignments for each option. A simple numerical measure based on the distribution of the responses will be discussed in the next chapter.

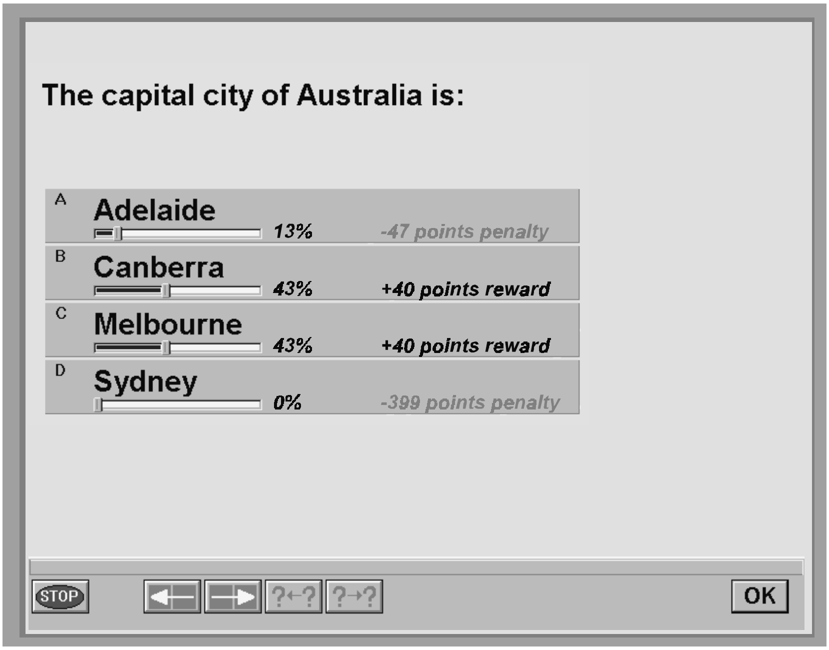
As a final indication of both item difficulty and item validity, the r.i.t. and r.i.r values can be calculated in the usual way, i.e. as the correlation between item scores and overall test scores (r.i.t.) and between item scores and residual test scores (r.i.r.).

5 - Experimental results

During the first semester of their first year, all students at the electronics faculty of our polytechnic take a course in Computer Systems. In previous years there was one weekly lecture period for all students, followed by one period for discussing any questions in groups of approximately 25 students. In the autumn 2000 semester, the ‘question time’ period was used for Multiple Evaluation testing. Initially 73 students took part, but 13 dropped out of the course for various reasons.
In the first week, the principles of ME testing were explained and students practised with the aid of ‘dummy’ (general knowledge) tests. Halfway through the course and at the end of the semester one week was dedicated to evaluation of ME versus MC by means of a questionnaire, and one other week was used for other work. In the course of the semester, this left 10 weeks for actual ME tests.

Based on earlier experience with 4-option Multiple Choice, it is standard practice in our faculty to set 10 problems per 50-minute period. (Standard MC examinations are scheduled in two or even four periods, for a total of 20 or 40 items.) However, we also wished to compare two possible implementations of ME testing: 4-option items using a computer and 2-option (True/False) pencil&paper questions. We therefore decided on two very short tests: five 4-option items by computer and ten 2-option items on paper.

For the computer tests we used TestBet [26]. Each item is presented on screen, with a ‘percentage-slider’ for each alternative answer as illustrated in Figure 11. As the student adjusts the percentage for each alternative, the percentages for the remaining alternatives are adjusted automatically to maintain a total of 100% distributed over all options. If it becomes necessary to readjust earlier settings, this is done proportionally so that the relative probabilities are maintained. Beside each bar the program displays the number of points the student will gain or lose if that particular alternative is the correct one. Once the student has confirmed his personal probability settings, the correct answer is highlighted. As the test progresses feedback is given regarding the student's realism, and the test concludes with a display of all results and conclusions.



The capital city of Australia is:

A Adelaide 13% -47 points penalty

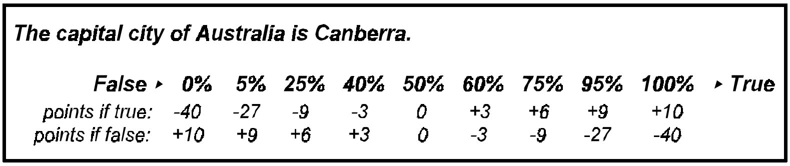
B Canberra 43% +40 points reward

C Melbourne 43% +40 points reward

D Sydney 0% -399 points penalty

Figure 11. Layout of TestBet screen.

For pencil&paper tests in general, the students themselves must ensure that the total of all probability assignments equals 100% and can then determine the corresponding reward or penalty points from a look-up table. This is obviously an additional and undesirable complication. There is one exception, however. When we use k = 2 items (true/false), setting the percentage for one option automatically determines the remaining percentage for the other. If we then limit the possible assignments to several discrete values, the look-up table can be included below each percentage as illustrated in Figure 12. In practice, we noticed that several students actually circled the points shown in the look-up table instead of the corresponding percentages - illustrating quite clearly on what they were basing their response!



The capital city of Australia is Canberra.

	False ▸	0%	5%	25%	40%	50%	60%	75%	95%	100%	▸ True
points if true:		-40	-27	-9	-3	0	+3	+6	+9	+10	
points if false:		+10	+9	+6	+3	0	-3	-9	-27	-40	

Figure 12. Layout of pencil&paper True/False item, with look-up table.

It is not possible to give automatic feedback either during or even immediately after the pencil&paper tests, obviously. The tests were scored the same day, however, and students were informed of all results (ME grades and realism scores for the computer and for the p&p tests) before taking the next week's tests. General information was also given to the group as a whole, concerning both the actual test questions and the average probability assignments for each option.

5.1 Scoring rules

The tests were scored according to the rules described earlier. To summarize:

- basic ME item scores were calculated according to the logarithmic rule, using the tolerance parameters t = 0.001 for computer (k=4) and t = 0.0362 for p&p tests (k=2). This sets the maximum penalty per item to −4 (with a maximum reward of +1). The (overall) realism score A was also computed.
- two realism-corrected ME scores were computed: both tolerance-adjusted and response-adjusted. Two corresponding overall ME test scores were calculated as the average item score times 10, resulting in a range from –40 to +10. The higher of these two was taken as final ME test score.
- the resulting ME test score (t-adjusted or r-adjusted) was converted to our standard 1 .. 10 point scale to obtain ME test grades.
- two hypothetical MC scores were calculated. For MC1 an item scored +1 if the highest probability assignment corresponded to the correct option, otherwise the item score was 0. MC2 used a sliding scale: +1 (or 0) if a clear choice was made for the correct (incorrect) option, and a score equal to the probability assignment r in the dubious mid-range. The test grades were the average item score times 9 plus 1, resulting in a range from 1 to 10.

All students were made aware of their (weekly) realism scores and final ME grades for the computer and p&p tests. Interested students could also view their basic ME and hypothetical MC1 and MC2 grades.

5.2 Item evaluation

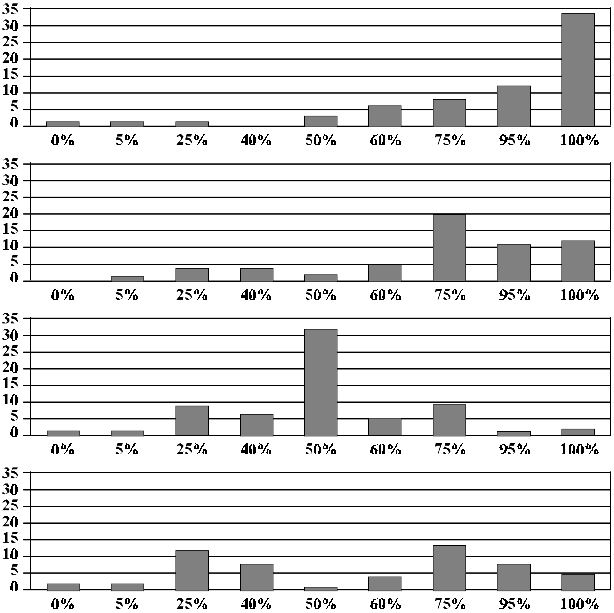
The first step in the test evaluation was an evaluation of the results for each item. The traditional r.i.t. and r.i.r. values were calculated, and the probability assignments were analysed in several ways. To facilitate this analysis, the computer results were first grouped into nine response percentage categories in such a way that the item scores in each group correspond approximately to those for the nine discrete response options in the p&p tests (see Table 4).

It should be noted that the values given in this Table are *response* percentages, and depend on our choice of tolerance parameter **t**. The *true* probability ranges can be calculated from the ME scores; the expected scores for two adjacent ranges should be identical for the borderline value of p. Between ranges #1 and #2, for instance, the *response* borderline is 2%. The *true* probability borderline, given the corresponding **k**=2 ME scores (–40 / +10 and –27 / +9), is calculated as:
$-40 \times p + 10 \times (1-p) = -27 \times p + 9 \times (1-p)$, so $p \approx 7\%$.
The **k**=4 ranges shown are based on the ME item scores for a given response.

range	#1	#2	#3	#4	#5	#6	#7	#8	#9
k=2, (%)	0 *(0-2)*	5 *(2-13)*	25 *(13-38)*	40 *(38-50)*	50 *(50)*	60 *(50-62)*	75 *(62-87)*	95 *(87-98)*	100 *(98-100)*
k=4, (%)	0-1	1-7	7-15	15-20	20-30	30-40	40-60	60-95	95-100

*Table 4. Probability assignment ranges for the computer tests (**k**=4). The item scores in these ranges correspond approximately to those for the corresponding discrete percentages in the p&p test (**k**=2).*

The number of probability assignments in each range are referred to as $n_1 \ldots n_9$. Even the simple expedient of plotting $n_1 \ldots n_9$ in a bar graph provides useful information. As illustrated in Figure 13 for some True/False items (all 'True'), the difference between 'easy', 'difficult', '*too* difficult' and 'dubious' items is immediately apparent. Easy items show a marked preference for the end-of-scale choices (100% or 95%, if the item is correct), whereas there is a much wider spread for 'difficult' items. When items are too difficult, the responses tend to cluster around 1/**k** (50% for **k**=2). Dubious items, on the other hand, lead to a 'camel's hump': some students clearly opt for 'True', whereas almost as many students consider the statement 'False'. It is interesting to note that this distinction would not be apparent in traditional Multiple Choice tests, where answers would be distributed approximately 50/50 in both cases.



$r\%_{av}$	diff(i)	r.i.t.	r.i.r.
86	2	0.80	0.75

$r\%_{av}$	diff(i)	r.i.t.	r.i.r.
75	3	0.60	0.50

$r\%_{av}$	diff(i)	r.i.t.	r.i.r.
51	4	0.52	0.38

$r\%_{av}$	diff(i)	r.i.t.	r.i.r.
57	3	0.45	0.27

Figure 13. Item analysis for 'easy', 'difficult', 'too difficult' and 'dubious' ***k****=2 items. (Note that each of these items was 'true'; for 'false' items the results would be mirrored horizontally.)*

These bar graphs are important for a first analysis of the item in general terms (‘easy’, ‘difficult’ or ‘dubious’), but we also used two numerical measures:

- The average probability assignments, $r\%_{av}$, give a first indication of both consensus and difficulty. As illustrated in Figure 13, the high percentages (86% and 75%) on the first two items correspond to ‘fairly easy’ and ‘difficult’; they also indicate that students in general agree that these items are ‘True’.

- A further measure of ‘item difficulty’, diff(i), is estimated as a weighted average of the number of responses in each range:
  $$\text{diff(i)} = [5\times(n_5) + 4\times(n_4+n_6) + 3\times(n_3+n_7) + 2\times(n_2+n_8) + 1\times(n_1+n_9)] / n \qquad (5.20)$$
  The result was rounded to a whole number from 1 to 5.
  If all n students select the central range (1/**k**, ‘don’t know’), diff(i) = 5; whereas if all n students select the outer ranges (0% = certainly incorrect or 100% = certainly correct), then diff(i) = 1.

Note that neither of these simple measures distinguishes between ‘difficult’ and ‘dubious’ items. However, this distinction is immediately apparent in the bar graph.

Figure 14 shows the set of bar graphs for some 4-option questions, where the results for each option are plotted separately. (For clarity, the options are ordered so that ‘A’ is the correct answer in all cases.) The ‘item difficulty’ is calculated in the same way as for **k**=2, but using the **k**=4 percentage ranges given in Table 4.

As a further indication of both consensus and difficulty, the average ME scores are given for each option. A good item should show a slightly positive average ME score for the correct option and slightly negative scores (ideally: equal) for the incorrect responses. Bearing in mind that the most negative score possible is −4 (with the tolerance parameter used), scores more negative than −2 or so indicate that the option was too obviously incorrect.

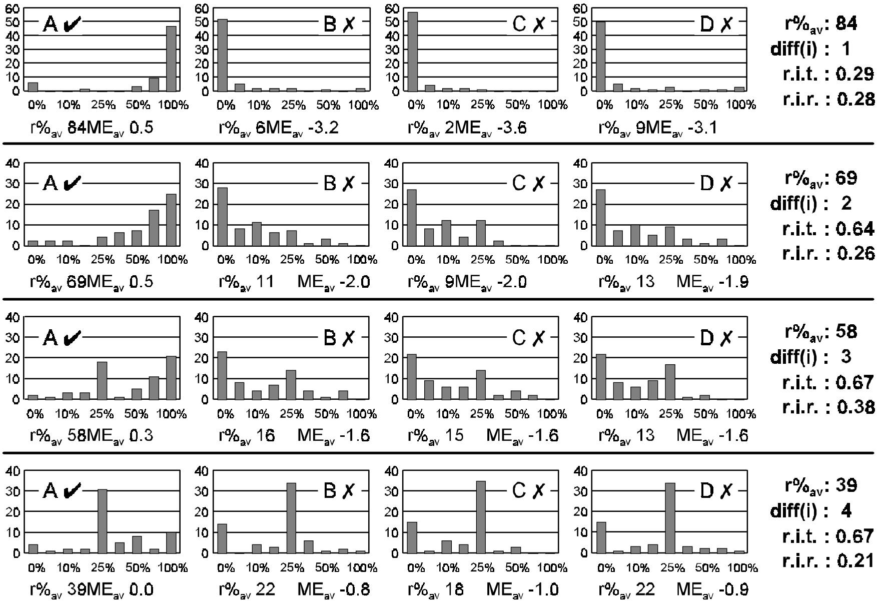


Figure 14. Item analysis for 'too easy', 'easy', 'difficult' and 'too difficult' k=4 items. (Option A is the correct answer in all cases.)

5.3 Test evaluation

For some of the paper&pencil tests, up to three items were removed (based on item analysis) and the estimates of test reliability were based on the remainder. The computer tests (consisting of only five items) were left intact.

The final scores for the ME tests are based on corrected-for-realism responses, as described in section 4.4, and therefore the test reliability estimates are based on these corrected responses.

For the MC1 and MC2 test results, Cronbach's alpha was calculated as a measure of test reliability and the required test length (**M**0.75) was estimated for $\alpha = 0.75$. Four estimates were used for the ME results, as derived in Chapter 4:

- Cronbach's alpha, calculated from the variance of the observed item scores and the variance of the observed final scores. These item scores are calculated from the response for the correct alternative r_c, with correction for realism, according to the logarithmic scoring rule. The variance σ_i^2 is calculated for each item and the results are summed to obtain the overall item variance.

- rho1, rho2 and rho3, calculated from the expected score *E(S)* (Formula 4.14), item variance *E*[Var(*S*)] (Formula 4.15) and/or the variance of the observed final scores σ^2_X , using the test reliability estimates (4.16), (4.17) and (4.18):

  $\boldsymbol{\rho 1} = \mathrm{Var}[E(S)] / (\mathrm{Var}[E(S)] + E[\mathrm{Var}(S)])$
  $\boldsymbol{\rho 2} = 1 - E[\mathrm{Var}(S)] / \sigma^2_X$
  $\boldsymbol{\rho 3} = \mathrm{Var}[E(S)] / \sigma^2_X$

The Spearman-Brown estimate of the required test length, **M**0.75, was based on rho1. The results for all tests are given in Table 5.

The (estimated) ‘hard’ Multiple Choice (MC1) results for **k**=4 correspond quite well with what we expected. Many of these items were derived from items in a well-calibrated item bank and, as stated earlier, current practice is to use 20 or 40 items per test. This test length is confirmed by the estimated test length for $\alpha = 0.75$. With a few exceptions, 13 to 24 items are indicated; averaged over all tests, 34 items are required for $\alpha = 0.75$. However, it should be noted that these estimated test lengths are based on the results for only five items.
We have very little experience with **k**=2 Multiple Choice items, but we had expected estimated test lengths of 60 to 80 items. The results proved surprisingly good therefore: averaged over the total period a test length of some 41 items is indicated. One possible explanation was provided by the students themselves. They remarked that they considered each item in these ME tests with greater care than they would have done for traditional MC, and our ‘MC’ results were derived from their ME responses.

The estimated uncertainty-scale MC2 results show an improvement for **k**=4, but they seem similar to or sometimes even worse than the MC1 results for **k**=2. As explained earlier, these scores are derived by setting each item score to 1 or 0 if there is a clear ‘true’ or ‘false’ preference for the correct option and using the actual response as item score in the (dubious) mid-range. Given the values for alpha, it would appear wise to reserve judgement on these MC2 scores.

Computer (k=4, 5 items)		**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**
alpha -	MC1	0.15	0.38	0.54	0.39	0.38	0.45	0.16	0.52	0.48	0.28
	MC2	0.23	0.56	0.61	0.33	0.48	0.62	0.19	0.58	0.67	0.34
	αME	0.00	0.43	0.56	0.40	0.44	0.52	0.31	0.57	0.67	0.42
rho -	ρ1ME	0.58	0.74	0.74	0.67	0.69	0.67	0.66	0.72	0.75	0.68
	ρ2ME	0.34	0.68	0.69	0.55	0.58	0.57	0.55	0.65	0.70	0.56
	ρ3ME	0.92	0.88	0.89	0.92	0.92	0.85	0.86	0.91	0.90	0.96
M0.75 -	MC1	83	25	13	24	24	19	82	14	16	39
	MC2	51	12	10	27	16	10	64	11	8	30
	ρ1ME	11	6	6	8	7	8	8	6	5	7
P&P (k=2, ≤10 items)		**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
	items removed	1	1	0	1	2	2	3	0	1	0
alpha -	MC1	0.65	0.52	0.62	0.37	0.23	0.39	0.26	0.63	0.43	0.30
	MC2	0.64	0.42	0.82	0.42	0.32	0.33	0.41	0.73	0.54	0.51
	αME	0.79	0.59	0.89	0.63	0.58	0.61	0.69	0.80	0.68	0.81
rho -	ρ1ME	0.80	0.59	0.90	0.70	0.66	0.68	0.63	0.87	0.66	0.81
	ρ2ME	0.83	0.64	0.90	0.70	0.67	0.69	0.65	0.86	0.63	0.82
	ρ3ME	0.61	0.51	0.84	0.70	0.62	0.65	0.60	0.91	0.71	0.79
M0.75 -	MC1	15	25	19	47	83	39	59	18	36	69
	MC2	16	38	7	37	51	48	31	11	24	29
	ρ1ME	7	20	4	12	13	12	13	5	15	8

*Table 5. Estimated test reliabilities (alpha and rho) and estimated test lengths for test reliability ≥0.75 (**M**0.75 = $\mathbf{m} \times \mathbf{L}_{0.75}$).*
For the Pencil&Paper tests up to three items were removed on the basis of item analysis.

In general, the Multiple Evaluation (ME) test reliability estimates based on Cronbach's alpha are quite good for the paper&pencil **k**=2 tests. For **k**=4 they seem rather disappointing, however: marginally better than MC1 and similar to or worse than MC2. The first test even appears to be a complete disaster.
Based on the simulations and analysis described in Section 4.8 this was to be expected for a **k**=4 test, however. Most students tend towards overestimation, certainly for this first test, and after correction-for-realism this leads to low values for alpha and ρ2, and a high value for ρ3. All **k**=4 tests (C1 .. C10) show the same effect. For the **k**=2 tests (P1 .. P10) the various test reliability estimates correspond more closely.
Given our theoretical analysis, we had decided to take **ρ1** as a first estimate for reliability and test length. The theory also predicted a marked improvement over MC, but even so the actual improvement far exceeded our expectations. An average of just seven items for a sufficiently valid **k**=4 test and ten items for a sufficiently valid **k**=2 test seem too good to be true.

In a sense, these results *are* indeed too good to be true. In such short tests, each and every item must be valid – there is little or no room for 'cleaning up' *after* the test. With traditional Multiple Choice many more items are required, both to cover the subject matter and to enable us to compensate statistically for 'gambling'. When using Multiple Evaluation, the problem of 'gambling' disappears and we should only require sufficient items to cover the subject matter. In actual practice we will often need more than this minimum number of items, to enable us to compensate for our own errors when setting the questions. For the ten-item **k**=2 tests we could afford the luxury of removing one or two items, thereby improving the reliability of the test results. For the five-item **k**=4 tests this was considered impractical - fortunately it also proved unnecessary.

5.4 Overall test evaluation

So far, we have only considered the individual (weekly) tests. Given the results for ten of each type of test (P&P **k**=2 and computer **k**=4) in the course of the semester, we can also consider various longitudinal evaluations. For individual students we can estimate their learning curves for ‘realism’, and compare their (average) final grades based on ME and MC1 scoring. Given a total of 100 True/False and 50 **k**=4 items we can perform (statistical) split-half comparisons in various ways.

As is so often the case when large quantities of data are available, the possibilities for evaluation and calculation are almost endless and we must restrict ourselves to those that may be expected to answer our most burning questions.

The main hypothesis is that Multiple Evaluation should prove a valuable test method and a major improvement over Multiple Choice. To some extent this is already borne out by the test results for each session given in Table 5, but for further confirmation we can compare the weekly test results for each student to his overall average.

A second point worthy of consideration is the score conversion rule, intended to convert the basic ME scores to ME grades on our standard 10-point scale. If we assume that average final scores based on all (estimated) MC results will give an acceptable approximation of traditional scoring results, we may hope that the average final score based on the ME grades is quite similar in most cases.

As a final point we can evaluate the individual ‘realism’ scores. Earlier research [21] indicates that most students tend to overestimate initially, but soon learn to give more realistic probability assignments.

Before attempting these evaluations, however, we must decide which results to allow. Of the 73 students, only 35 actually completed all tests. These will be referred to as Group 1. For a further 26 students (Group 2) *most* results are available, but not all (for various reasons: late starters, illness, even computer error). Finally, Group 3 consists of 12 students who followed only part of the course. Most of these are ‘early drop-outs’, so that their results during the first few weeks can be validly compared with those of the other two groups.

To present a total picture we will evaluate the results for all students, after dividing them into the three groups described above. This has the added advantage that we can compare overall group results.

When comparing the weekly test results to the overall averages, two further factors must be considered: differences in test difficulty and student-attitude changes in the course of the semester. We can compensate for test difficulty to some extent, either on the basis of the estimated item difficulty or on the basis of the average test score for all students - or even a combination of both these indicators. Compensation for changing attitudes (e.g. ‘working harder’ or ‘giving up’) is far more difficult to accomplish without corrupting the very data that we wish to evaluate, so no attempt was made in this direction.

If we assume that most students are fairly consistent, their weekly test scores should be fairly close to their overall average result. Compensation for test difficulty should improve this match. To test this hypothesis the variance of the test scores was calculated for each student, and the results for the whole group were then compared. Given the fairly small number of tests, it is to be expected that students who missed one or even several of these tests will influence the results noticeably. For this reason Group 1 was also considered separately, but the results proved quite similar to those for the whole group. The results are given in Table 6 (Computer, **k**=4) and Table 7 (P&P, **k**=2).

Computer (k=4)		Variance of uncorrected weekly test grades			Variance of test grades after correction		
		min.	average	max.	min.	average	max.
All students:	MC1	0.80	4.34	13.44	0.00	2.78	8.02
	MC2	0.11	3.56	9.71	0.02	2.23	5.73
	ME	0.02	2.61	6.18	0.10	1.76	4.50
Group 1 only:	MC1	0.80	4.03	13.44	0.62	3.05	7.57
	MC2	0.86	3.29	9.71	0.33	2.11	4.37
	ME	0.96	2.64	6.18	0.55	1.98	4.50

*Table 6. Minimum, average and maximum variance of student grades on the weekly **k**=4 tests, before and after correction for ‘difficulty’.*

P&P (k=2)		Variance of uncorrected weekly test grades			Variance of test grades after correction		
		min.	average	max.	min.	average	max.
All students:	MC1	0.00	3.02	11.60	0.00	1.97	5.92
	MC2	0.00	1.81	5.04	0.07	1.15	3.76
	ME	0.01	1.88	4.93	0.07	1.24	3.12
Group 1 only:	MC1	1.44	3.28	6.44	0.64	2.06	4.80
	MC2	0.75	2.06	5.04	0.41	1.22	3.76
	ME	0.52	2.10	3.73	0.30	1.26	2.89

*Table 7. Minimum, average and maximum variance of student grades on the weekly pencil&paper (**k**=2) tests, before and after correction.*

Multiple Evaluation is a clear winner when compared to the estimated ‘hard’ Multiple Choice results (MC1). The ‘uncertainty-scale’ MC results (MC2) come in as a good second. However, it must be stressed that these MC2 scores are based in part on the probability assignments elicited on the basis of the logarithmic ME scoring rule, and as such lack the hard dichotomic character of a true Multiple Choice test.

Having established that the weekly ME grades give a more accurate indication of student performance than the (estimated) MC scores when compared to the corresponding overall averages, we must now consider whether these results may be used as final grades. Taking the overall MC scores as an indicator of current scoring practice, the overall ME grades should correspond numerically. If this proves to be the case, the score conversion rule can be considered satisfactory.

As a first step, the correlations between the average final grades (ME, MC1 and MC2) for all students are compared in Table 8. For both the **k** = 4 and the **k** = 2 tests these correlations are satisfactorily high. The correlations between the final grades for **k** = 4 and **k** = 2 are also given in Table 8, at the right, and these are distinctly smaller. For example, the correlation between the ME **k** = 4 and **k** = 2 grades is only 0.448. We will return to this point later, in section 5.6.

k = 4	
ME - MC1:	0.93
ME - MC2:	0.97
MC1 - MC2:	0.94

k = 2	
ME - MC1:	0.88
ME - MC2:	0.96
MC1 - MC2:	0.87

k = 4 vs. **k** = 2	
ME :	0.448
MC1:	0.587
MC2:	0.541

*Table 8. Correlations between the average grades for all students, for the various scoring methods, for the **k** = 4 and **k** = 2 tests.*

*Also shown, at the right, are the correlations between the final **k** = 4 and **k** = 2 results for the three scoring rules.*

The numerical correspondence between the final grades is plotted in Figure 15. As these plots clearly illustrate, the ME grades correspond quite well to the (estimated) MC results. The score conversion rule already provides a satisfactory mapping of the basic ME scores onto our standard 10-point scale.

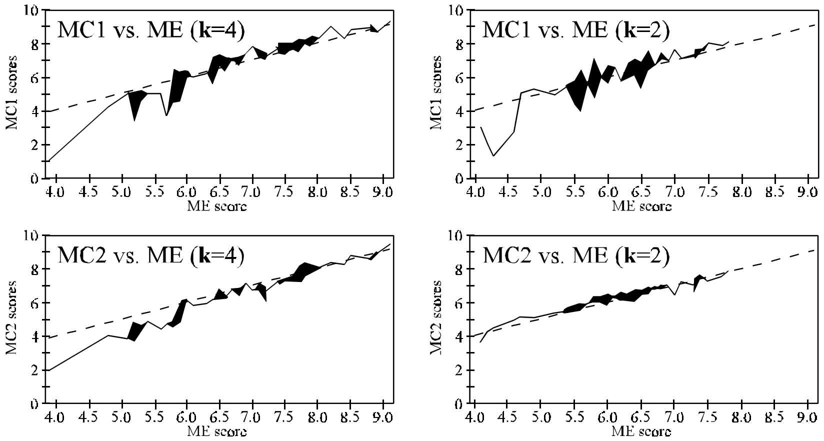


Figure 15. Final 'hard' MC1 and 'uncertainty-scale' MC2 scores, plotted against the ME grades obtained using the score conversion rule.

The dotted line corresponds to a perfect match (equality), and the dark areas correspond to a spread in the MC scores.

There is still room for improvement, however: at the low end of the scale the conversion rule is perhaps too tolerant. Where the ME grades are falling from 6 down to 4, some of the corresponding MC scores are falling to lows of 2 and 1. This possibility was mentioned earlier, in section 4.2, where we showed that 'zero knowledge' (ME score = 0) would lead to a final ME grade of 4.38.

This lenient score conversion has the disadvantage that (too) many students have final grades just below our pass/fail cutting score (5.5), and may be led to believe that they already have 'almost sufficient' knowledge and proficiency. A more severe conversion rule for lower scores could increase their motivation towards further study. On the other hand, the rule must leave sufficient 'bottom clearance' *below* the zero-knowledge score to serve warning to overconfident students with severe misconceptions (who will receive negative ME scores). Such a rule will be discussed in Chapter 6.

It should be noted that the score conversion rule affects the *overall* test result, especially for low scores; the tolerance parameter **t** sets the severity *per item*, especially for low probability assignments. Although these are two different aspects of 'severity', we may also expect that a more severe score conversion rule could provide an added incentive towards more realistic probability assignments. A student with insufficient knowledge (leading to a low overall test score) should not overestimate his capabilities (risking severe penalties for individual items).

As a final point we can evaluate the individual 'realism' scores. Earlier research (in particular by Dirkzwager [21]) indicates that most students tend to overestimate initially, but soon learn to give more realistic probability assignments. On the other hand, Shuford [73] warns that several factors may have a detrimental effect on this. For instance, a specific 'cutting' score (as, in our case, 'pass' for a score ≥ 5.5) may result in less motivation towards accurate probability assignments. Many students tend to aim for a 'pass' that is achieved with a minimum of effort, and answering a test on the premiss that "6 is enough" gives little incentive to optimize the score.

Furthermore, the score conversion rule takes some of the ‘sting’ out of Multiple Evaluation. If just one wildly incorrect answer results in a *negative* score, you are inclined to think twice next time. If it only reduces the overall grade by one or at most two points out of ten, that incentive is greatly diminished.

Toward the end of the semester, in our case, a further factor enters the equation. Students were aware that their final grade would be based on the average results of all tests, and many were already in the comfortable position that even a zero score for the last few tests would leave them with an average ‘pass’. This could lead to a lack of realism towards the end, either by students avoiding risks (underestimating) or by becoming less careful (overestimating). On the other hand, students with a low average score might be inclined to take greater risks in the hope of improving their average score on the basis of ‘lucky hits’.

In summary, we may expect that students ‘soon learn to give more realistic probability assignments’, but might become progressively more ‘calculating’ in the course of the semester with the result that they become *less* realistic toward the end.

In practice, the results proved rather better than expected. The results for all students were plotted, and some typical examples are shown in Figure 16. Using the realism measure described earlier, ‘perfect realism’ corresponds to the value 1; overestimating gives smaller or even negative values (*above* the centre-line in the plots), and underestimating gives larger values (*below* the centre-line). The box plots illustrate the results for all students, for the ten computer tests (**k**=4) and the ten p&p tests (**k**=2).

Some students did indeed overestimate initially, and a few then overcompensated wildly for the second test by seriously *under*estimating. However, most students soon learned to give quite realistic probability assignments. As the overall average plots show, the initial tendency to overestimate soon disappeared. There was no evidence to indicate that students became less realistic toward the end of the semester.

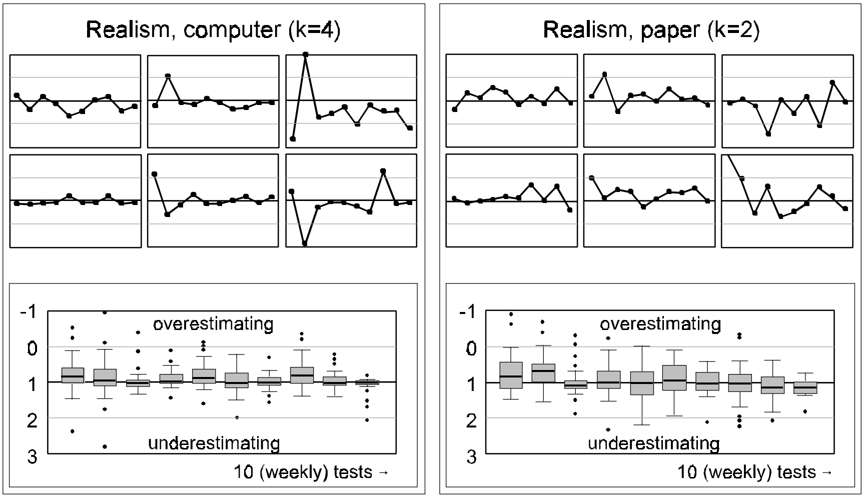


*Figure 16. Examples of 'realism' plots for individual students, both for the weekly computer (**k**=4) and for the P&P (**k**=2) tests.*
The box plots give the overall results for all students, in the course of the semester (10 weekly tests).

Compared to the computer tests, the P&P tests show a distinctly greater tendency towards overestimation. This may be partly due to the discrete scoring scale, but the results tend to indicate that there is also a residual 'Multiple Choice' effect. When in doubt, students are quite willing to select a mid-range percentage (between 25% and 75%), but when they feel 'fairly certain' they tend to favour the two extremes (0% or 100%). The '5%' and '95%' options are distinctly less popular.

The 'cutting score' also had a negative effect. Nearly all students took the computer test first and then, already knowing that score, took the P&P test. Some quite honestly stated that they put far less effort into that: "I got at least 8 out of 10 on the computer, so even 4 out of 10 on paper is enough".

Despite these effects, the P&P results are quite good (as illustrated in Table 5). Where computers are not (yet) available in sufficient numbers, pencil and paper is a valid option for Multiple Evaluation testing.

Nonetheless, computer testing is to be preferred for an entirely different reason: on-line feedback. Our students (who had the opportunity to compare both options) clearly indicate that they themselves consider this an important feature, sadly lacking in all types of pencil&paper tests.

5.5 Evaluation by the students

After the first quarter, that is after six weeks of testing, the students were asked to evaluate this system (interim evaluation, Figure 17). At the end of the semester they evaluated it a second time (final evaluation, Figure 18). The questions were presented in our standard format: introductory text which is to be considered as certainly true in italics, followed by the statement to be evaluated in plain text.

As a further experiment, the questionnaires were answered and scored according to a variation on Multiple Evaluation. This was based on a suggestion by Dirkzwager [25] which he called "BetVote", intended to train participants in a discussion to make good use of their estimate of the general consensus in a group.
As an extension of this concept, it seemed interesting to ask our (well-calibrated) students to respond to a questionnaire which was presented as a paper&pencil ME test. (To be consistent with his 'TestBet' and 'BetVote', this variation could be termed 'BetterSurvey'?) This system is explained in greater detail in Appendix 3, but the principle can be summarized briefly as follows:

- Students respond on the customary P&P **k**=2 scale. For each item, the average percentage score (over the whole group) determines whether the statement is considered 'True' or 'False', initially. Based on this 'key', ME scores for the total questionnaire are then calculated for all students.
- Based on these ME scores, the original responses (percentage assignments) of all students for all items are 'weighted': the response percentages for students with higher ME scores are shifted to more extreme values, whereas for low ME scores the responses are shifted closer to 50%. This has the effect that students whose responses approximate the overall opinion of the whole group (giving

  them a higher ME score) will be taken ‘more seriously’; radical dissidents have far less effect on the result.

  To stress this basic idea, all questions were phrased as “Most students feel that ...” instead of the more commonly implied “I feel that...”. Students were made clearly aware that these evaluations were intended as a measure of the general opinion of the whole group - not as a means to express personal grievances.

- New overall-average percentage scores are calculated on the basis of the weighted responses.

Figures 17 and 18 give the original (unweighted) average percentages and the (weighted) final results for each question, with a bar graph of the original responses (over the nine percentage options from 0% to 100%). The weighted results consistently give a clearer yes or no response to the questions.

The first statement in the interim evaluation, which we will refer to as Q1-1, was repeated verbatim as statement 7 in the final evaluation (Q2-7): “Most students agree that ME gives fairer and more accurate scores than MC”. Initially there was an almost exact 50/50 split between Agree and Disagree and the weighted responses showed 59% agreement. At the end of the course the general consensus had shifted further towards agreement: 59% unweighted and 69% weighted.
This response may well be due in part to the fact that most semester-end grades were satisfactorily high (partly for other reasons, as will become apparent). Understandably, a good score may be taken as an indication of a ‘fairer’ scoring system.

The student judgements on other aspects of ME versus MC are more clear-cut. Students definitely appreciate the fact that “ME does not force you to pick just *one* of the answer options” and allows you to give a more differentiated response (Q1-2: 69% ▸ 75%); the majority also feel that “ME forces you to consider all options more carefully than for MC” (Q2-6: 66% ▸ 75%). Most students “at first found it quite difficult to give a realistic estimate of their personal probabilities” (Q1-5: 82% ▸ 84%), but “soon learned to give realistic estimates” (Q1-6: 74% ▸ 79% and Q2-8: 76% ▸ 81%).

1 *The individual scores for the first six weeks are known.*
Most students agree that ME gives fairer and more accurate scores than MC.
51% ⇒ 59% : Marginal

2 *ME lets you score points even when in doubt.*
Most students like the fact that ME does not force you to pick just *one* of the answer options.
69% ⇒ 75% : Agree

3 *ME tests were taken both by computer and via P&P.*
Most students prefer the computer ME tests to those by means of pencil&paper.
78% ⇒ 80% : Agree

4 *The P&P items were True/False; computer items were 4-choice, although True/False is possible.*
Most students prefer True/False items to 4-choice.
55% ⇒ 65% : Agree

5 *ME requires realistic probability assignments.*
Most students at first found it quite difficult to give a realistic estimate of their personal probabilities.
82% ⇒ 84% : Agree

6 *By now, 12 tests have been taken using ME rules.*
Most students soon learned to give realistic estimates of their personal probabilities.
74% ⇒ 79% : Agree

7 *The ME results include a score for 'realism'.*
Most students are only interested in their final grades, and consider all other information irrelevant.
67% ⇒ 71% : Agree

8 *The average of the weekly tests is the final score.*
Most students would prefer one final examination, instead of the weekly tests.
25% ⇒ 21% : Disagree

9 *For ME, you must know the extent of your knowledge.*
Before each weekly test, most students are already well-aware of what they do and what they don't know.
60% ⇒ 67% : Agree

10 *For ME, you must know the extent of your knowledge.*
After each weekly test, most students are far more aware of what they do and what they don't know.
68% ⇒ 74% : Agree

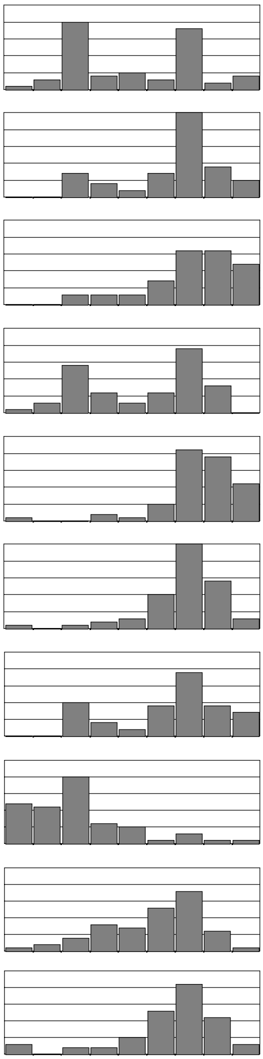
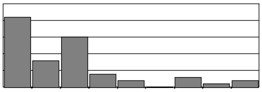
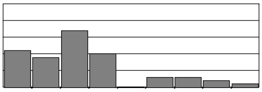


Figure 17. Student (interim) evaluation of Multiple Evaluation testing, showing both the average response percentage and the weighted percentage for each item. The bar graphs show the response in nine categories from 0% to 100%.

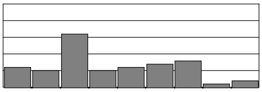
1 *The average of the weekly tests is the final score.*
Most students would prefer one final examination, instead of the weekly tests.
21% ⇒12% : Disagree



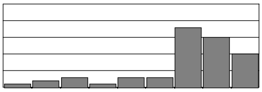
2 *For weekly tests you need to study regularly.*
Most students feel that they would learn just as much by intensive study just before a final examination.
28% ⇒25% : Disagree



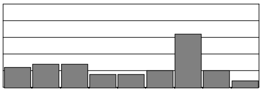
3 *The weekly tests replaced a regular lecture period.*
Most students feel that they would profit more from an extra lecture period than from the weekly tests.
40% ⇒32% : Disagree



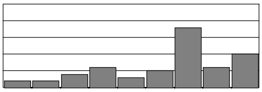
4 *ME tests were taken both by computer and via P&P.*
Most students prefer the computer ME tests to those by means of pencil&paper.
76% ⇒85% : Agree



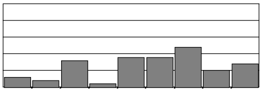
5 *Comparing True/False items to 4-choice:*
Most students prefer True/False items to 4-choice (for this subject).
49% ⇒36% : Disagree



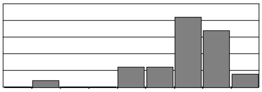
6 *ME lets you score points even when in doubt.*
Most students feel that ME forces you to consider all options more carefully than for MC.
66% ⇒75% : Agree



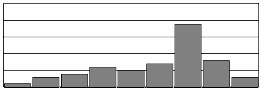
7 *The final grade (overall average score) is known.*
Most students agree that ME gives fairer and more accurate scores than MC.
59% ⇒69% : Agree



8 *At first, most students found it difficult to give realistic estimates of probabilities. Now, looking back:*
Most students soon learned to be ‘realistic’.
76% ⇒81% : Agree



9 *All students received both ‘test’ and ‘realism’ scores.*
Most students are only interested in their weekly grades, and consider all other information irrelevant.
63% ⇒71% : Agree



10 *After each weekly test, most students are far more aware of what they know than before the test.*
This is true for *any* test: for ME no more than for MC.
52% ⇒63% : Agree

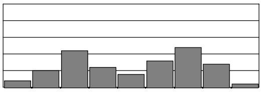


Figure 18. Student (final) evaluation of Multiple Evaluation testing, showing both the average response percentage and the weighted percentage for each item. The bar graphs show the response in nine categories from 0% to 100%.

Rather disappointingly, students indicate that they are hardly interested in their realism scores as such: "Most students are only interested in their (..) grades and consider all other information irrelevant" (Q1-7: 67% ► 71% and Q2-9: 63% ► 71%).

Tests are a learning experience: "*Before* each test, students are well-aware of what they do and what they don't know" (Q1-9: 60% ► 67%), but "*After* each test, students are far more aware of what they do and what they don't know" (Q1-10: 68% ► 74%). However, they don't restrict this observation to ME: "This is true for any test: for ME no more than for MC" (Q2-10: 52% ► 63%). Many students elaborated on this statement by adding the proviso: "provided immediate feedback is given during the test, as here for ME, so that you can re-consider what you did wrong". This may well explain their clear preference for computer tests as opposed to the p&p version (Q1-3 and Q2-4).

There is no such clear preference for the type of item. Initially they tended to prefer True/False to 4-choice (Q1-4), but the final evaluation indicates a preference for 4-choice (Q2-5). Note, however, that this choice may be based on their preference for computer testing despite the (more specific) phrasing of the original question: "... irrespective of whether on paper or via the computer".

The students clearly voted for weekly tests as a replacement for one final examination (Q1-8 and Q2-1), also as an encouragement to learning (Q2-2) even when these tests replace a weekly lecture period (Q2-3). This is not peculiar to ME, of course, but it is an important factor to consider for future curriculum developments. Improved learning behaviour is reflected in higher scores, and this group definitely did better than their predecessors, as illustrated in Table 9.

In summary: the students consider ME a fairer and better system than MC, and feel that ME forces them to consider the questions and answer options more carefully. They also appreciate the immediate feedback and the weekly tests as aids to learning - irrespective of whether ME or MC scoring is used.

semester	% pass	% fail	scores (scale 1 .. 10)		
			lowest	average	highest
Autumn 1998 (MC)	82	18	3.9	6.7	10.0
Spring 1999 (MC)	43	57	2.6	5.0	7.8
Autumn 1999 (MC)	71	29	1.0	5.9	8.9
Spring 2000 (MC)	43	57	1.1	4.8	8.9
Autumn 2000 (ME)	**87**	**13**	**4.2**	**6.4**	**7.9**

Table 9. Results after weekly testing (Autumn 2000) compared to earlier results with overall MC tests at the end of the course.
Note 1: the lowest score for Autumn 2000, 4.2, can be attributed in part to the less severe score conversion rule applied for ME tests.
Note 2: the Autumn 1998 and 1999 groups included many students with prior knowledge of this subject; in both Spring groups and in the Autumn 2000 group these students were exempted.

5.6 Other comparative results: validity

While summarising earlier research (in Chapter 1) we stated: “On the ‘validity’ aspect, the jury is still out. When rated on the basis of correlation with some criterion, some results are better and others are worse.”

One of the major difficulties is the selection of a suitable criterion, certainly when experiments are conducted in a regular educational environment. ‘Other’ tests may be used (such as prior or parallel MC tests, short-answer tests or semester-end grades), but then we may question the validity of the criterion: any difference between ME results and the criterion may be interpreted as an indication that ME is either better or worse. Alternatively we can look for a more general indication of the student’s proficiency, such as a subjective rating by supervisors during practical work.

For our research we had access to all semester-end grades and subjective ratings for practical work. Unfortunately the latter proved insufficiently graduated, with only three possible results: ‘sufficient’ (for the majority of the students), ‘good’ (for 15%) or ‘insufficient’ (for 3%). Since our subject is ‘Computer Systems’, we decided to compare the results for the following subjects:

- **Computer Systems**, final grade;
- Software Programming (results based on structured programming assignments);
- Digital Electronics (results based on **k**=4 MC test);
- Delphi Programming (results based on structured programming assignments);
- Mathematics (results based on **k**=4 MC tests and take-home assignments);
- Electrical Network theory (results based on **k**=4 MC tests);
- Analogue Electronics (results based on **k**=4 MC test);
- Power and Motion Systems (results based on **k**=4 MC test);

The first four of these are ‘digital’ subjects, the other four are ‘analogue’.

Using principal component analysis we found two distinct components, as Table 10 shows. The first component seems to correspond to ‘overall proficiency’, and the second makes a clear distinction between the ‘digital’ and ‘analogue’ subjects. After Varimax rotation this distinction becomes even more apparent, with one clearly ‘analogue’ and one clearly ‘digital’ component.

It is interesting to note that Computer Systems is the most clearly ‘non-analogue’ subject and Electrical Network Theory is the most obviously ‘non-digital’, and precisely these two subjects have the lowest loading on the first unrotated component which we have dubbed ‘overall proficiency’.

Judged on the basis of this principal component analysis, the final grades for ‘Computer Systems’ are valid. However, the doubts raised against Multiple Evaluation testing in general are based on the assumption that a personality factor (‘realism’ or ‘self-confidence’) may contaminate the results.

Principal components:
Component 1, explained variance: 3.425
Component 2, explained variance: 1.353
(Component 3, explained variance: 0.967)

	Component Matrix		Varimax Rotated Matrix	
	1	2	1	2
Computer Systems	0.454	0.574	–0.056	0.729
Software Programming	0.602	0.440	0.143	0.732
Digital Electronics	0.655	0.274	0.295	0.646
Delphi Programming	0.794	0.345	0.349	0.792
Mathematics, first course	0.675	–0.361	0.741	0.193
Network Theory	0.517	–0.562	0.761	0.062
Analogue Electronics	0.785	–0.371	0.828	0.260
Power and Motion	0.677	–0.226	0.651	0.294

Table 10. Results of principal component analysis. The first four subjects are 'digital', the other four are 'analogue'.

To test this hypothesis a second principal component analysis was performed, using the following Multiple Evaluation final grades:

- final test grade paper&pencil (10-item, **k** = 2)
- final test grade computer-based (5-item, **k** = 4)
- final test grade overall (unweighted average of the above results)
- final score for realism, paper&pencil (10-item, **k** = 2)
- final score for realism, computer-based (5-item, **k** = 4)
- final score for realism, overall (unweighted average of the above results)

The results are given in Table 11.

Principal components:
Component 1, explained variance: 2.631
Component 2, explained variance: 1.609
(Component 3, explained variance: 1.109)
((Component 4, explained variance 0.390))

	Two component analysis				Three components, Rotated Matrix		
	Components		Rotated				
	1	2	1	2	1	2	3
P&P grade	0.798	0.065	0.652	0.464	0.425	0.734	0.184
Computer grade	0.394	0.829	0.009	0.914	0.919	–0.065	–0.010
Overall grade	0.624	0.680	0.188	0.904	0.896	0.193	0.112
P&P realism	0.612	–0.407	0.734	–0.004	–0.091	0.953	0.039
Computer realism	0.593	–0.267	0.646	0.007	0.089	–0.026	0.995
Overall realism	0.850	–0.467	0.969	0.003	0.011	0.589	0.794

Table 11. Principal component analysis: 'final grade' versus 'realism'.

If we restrict ourselves to a two-component analysis, the rotated component matrix clearly distinguishes between 'realism' (rotated component 1) and 'grade' (rotated component 2). However, the final grade for the pencil&paper tests has a high loading on *both* of these components.

It would appear that (lack of) realism has a greater influence on the P&P grades. This effect is further explored in the three-component rotated matrix. The first rotated component again clearly separates the final test grade from 'realism'. The second component corresponds to P&P realism *and* has a high loading for P&P final grade and overall realism; the third component includes computer realism and overall realism.

The mixed results for the paper&pencil tests may explain the slightly lower correlation between the computer and P&P final grades given in Table 8.
One possible explanation for these mixed results is the fact that the score for the computer test (taken first) may influence a student's attitude towards the paper&pencil test. As illustrated in Figure 16, there was a wider spread between overconfidence and underconfidence on these tests, and this will lower the test scores.

Alternatively, these results may be seen as an indication that the P&P version is more prone to suffer from a 'personality' effect than computer-based testing. This, in itself, would not be surprising: students seem more inclined to choose '0% or '100%' on a printed sheet than they are to move a selection bar on the computer screen to either of the extreme positions. This may be due in part to a residual 'Multiple Choice' effect: when confronted with questions on paper, some students tend to revert to 'True/False' responses. In this connection it is interesting to note that earlier comparative research was mainly conducted on the basis of printed forms.
Whatever the reason, some students appear to have been less careful when taking these P&P tests and we cannot rule out a possible 'personality' effect.

To some extent, this hypothesis is borne out by one further principal component analysis.
Table 12 gives the analysis results when realism scores are compared to other semester-end grades. The final grades for the two programming subjects are determined to a large extent by the care taken for software assignments: accuracy both in program design and implementation. For Digital Electronics the final grade is based on a traditional **k** = 4 multiple-choice test.

The first rotated component is weighted heavily towards P&P realism and both the programming subjects. A personality trait, hasty and insufficiently careful work, may be the common factor. The second rotated component clearly relates to 'overall proficiency' only, and the third corresponds to 'realism' for the computer-based tests.

Principal components:
Component 1, explained variance: 3.254
Component 2, explained variance: 1.208
Component 3, explained variance: 1.036
(Component 4, explained variance 0.713)

	Component Matrix			Rotated Matrix		
	1	2	3	1	2	3
P&P score for realism	0.573	–0.350	0.637	0.924	–0.027	–0.036
Computer score, realism	0.623	–0.256	–0.738	0.018	0.133	0.990
Overall score for realism	0.848	–0.422	–0.164	0.605	0.085	0.742
Computer Systems	0.510	0.679	–0.026	0.068	0.845	0.051
Software Programming	0.757	–0.068	0.184	0.643	0.326	0.301
Digital Electronics	0.561	0.606	–0.033	0.124	0.809	0.111
Delphi Programming	0.819	0.102	0.154	0.602	0.505	0.296

Table 12. Principal component analysis: ‘realism’ scores versus final (semester-end) grades for ‘digital’ subjects.

Although the results of principal component analysis are not ideally clear-cut, we are inclined to read them as evidence that Multiple Evaluation testing leads to valid results.

This certainly appears to be the case for the computer-based tests, with correction for realism as incorporated in our research. For this and many other reasons, this appears to be the preferred testing environment. Unfortunately, lack of hardware facilities and inadequate software makes this difficult.

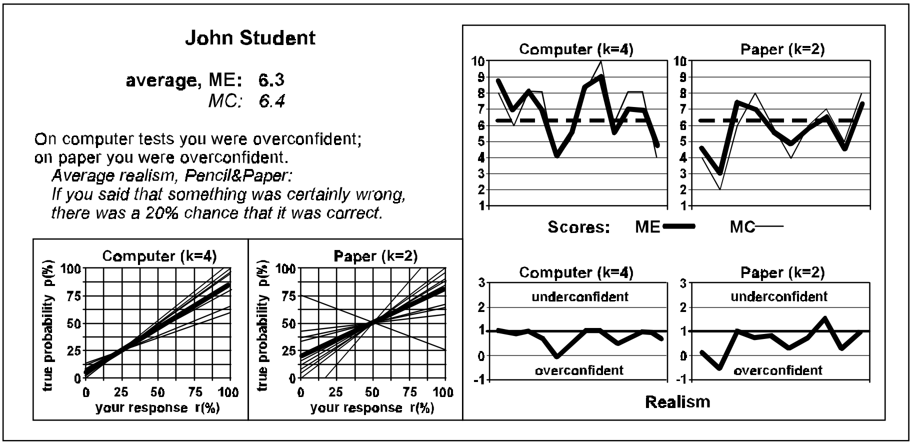
The P&P test results may suffer from a greater degree of ‘contamination’. However, in our ongoing research we are restricted to P&P tests (for the reasons given above) and the results tend to indicate that this is a viable poor-man’s alternative.

5.7 Feedback

It was noted earlier (specifically in section 4.3) that most students require some form of feedback concerning the realism of their probability assignments. This is certainly true during the initial learning phase.

In this study, the individual realism scores were discussed after each of the first tests and students were made clearly aware of the effect of their assignments on the scores. After this initial learning phase, all students still received their realism score for each test but further discussion was limited to those students who were still seriously over- or underestimating their capabilities.

Furthermore, at mid-term and at the end of the semester the students received a printed summary according to the format illustrated in Figure 19.



John Student

average, ME: 6.3
MC: 6.4

On computer tests you were overconfident;
on paper you were overconfident.
Average realism, Pencil&Paper:
If you said that something was certainly wrong, there was a 20% chance that it was correct.

Figure 19. Summary of individual results, handed out to the students.

This summary shows both the actual ME grades and (for comparison) the estimated MC scores at the top right; the overall averages are given below the student's name. In the interim and final evaluations, the majority of the students state that this is the only information that interests them - "all other information is irrelevant".

The other information all relates to the realism (A) of their probability assignments. The plots at the lower right show the realism scores for all tests. These were calculated by the least squares method, so that ‘perfect realism’ corresponds to a score of 1.

The plots at the lower left show the error introduced by lack of realism. Each line corresponds to one realism score, and the average realism score (over all tests) is shown as a bold line.

These lines are based on the relationship between response (r), realism (A) and true (personal) probability (p) given earlier (4.7):

$$p = \mathrm{A}\times r + (1\text{-}\mathrm{A})/\mathbf{k} \quad \text{(where } \mathbf{k} \text{ is the number of alternatives)}$$

For $\mathbf{k}$=2, a response r = 0% corresponds to an estimated true probability $p = (1 - \mathrm{A})/2$ and r = 100% corresponds to $p = \mathrm{A} + (1 - \mathrm{A})/2 = (1 + \mathrm{A})/2$. Similarly for $\mathbf{k}$=4, r = 0% corresponds to $p = (1 - \mathrm{A})/4$ and r = 100% corresponds to $p = (1 + 3\mathrm{A})/4$. Using these estimated values of p, straight lines in the plot show the (estimated) correlation between the probability assigned by the student (r) and the true personal probability (p). This information can be used by the student: “In future, when I feel like replying 0% it would probably be better to say 20%”.

Based on the average realism, further feedback is given as text. The first two lines give overall indications for both computer and paper tests:

- “... you were very overconfident” (for $\mathrm{A} < 0.5$)
- “... you were overconfident” (for $0.5 \leq \mathrm{A} < 0.8$)
- “... you were realistic (on average)” (for $0.8 \leq \mathrm{A} \leq 1.25$)
- “... you were underconfident” (for $1.25 < \mathrm{A} \leq 1.5$)
- “... you were very underconfident” (for $1.5 < \mathrm{A}$)

These ranges correspond to relationships between the student’s response and the true probability, as illustrated in Figure 20. On the overconfident side, a student is still considered realistic if a response (r) of 0% corresponds to a true probability (p) of up to 10% for $\mathbf{k}$=2 or up to 5% for $\mathbf{k}$=4 (and r = 100% corresponds to more than 90% and 85% respectively). Outside these ranges the student is considered “overconfident”, or even “very overconfident” if r = 0% corresponds to p > 25% ($\mathbf{k}$=2) or p > 12.5% ($\mathbf{k}$=4).

For underconfident students the borderline values for “you were realistic” are similar to those for overconfidence: a response (r) of up to 10% for **k**=2 or up to 5% for **k**=4 corresponds to a true probability (p) of 0% (and a response r greater than 90% and 85% respectively correspond to p = 100%).

The step to “very underconfident” is made slightly sooner, if r > 17% (**k**=2) or r > 8% (**k**=4) correspond to p = 0%, so that r < 83% (**k**=2) or r < 75% (**k**=4) correspond to p = 100%. These stricter limits were based on practical experience: few students tended to be underconfident, and a student who scored outside this limit could definitely be considered an exception - hence “you were *very* underconfident”.

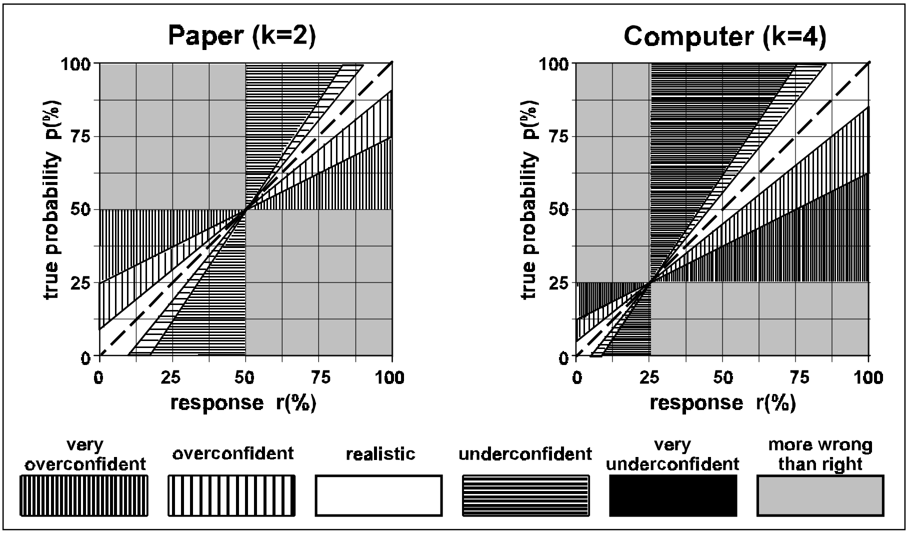


*Figure 20. Verbal feedback, from “you were very overconfident” to “you were very underconfident”, is based on the ranges shown above. For example: a student who responds for the **k**=2 test with 100% when the true probability is less than 75% and responds with 0% when the true probability is more than 25% is considered “very overconfident”. A student is considered realistic for **k**=2 if his most extreme responses are within 10% of the mark.*

The poorer of the two realism scores (for **k**=2 and **k**=4) was also presented as an explicit warning:

- "If you said that something was certainly wrong,
  there was a .. % chance that it was correct" (for overestimators)
- "If you said that something had a .. % chance of being right,
  then it was almost certainly correct" (for underestimators)

The actual percentages are derived from the formula given above.

Although students claimed that they consider this information "irrelevant", it should be noted that they at least appreciate the feedback. Several students who were not present when these slips were handed out asked for them later; the only remaining feedback notes at the end of the semester corresponded to students who had already left the course.

6 - Discussion

The first experimental results are quite promising. For practical applications in certifying (summative) testing a few major points must still be considered, however:

- the *decision accuracy*, especially for results close to the cutting score: how accurately are we discriminating between ‘pass’ and ‘fail’?
- the *score conversion rule* required to map the raw ME scores onto a standard scale. Certainly during a transition period, it is desirable for the final ME grades to be similar to traditional scores: their subjective ‘feel’ and numerical ‘weight’ should be sufficiently similar to those of conventional scores.
- the *error of measurement* as it reflects on the numerical value of the grades. This error is determined both by the intrinsic inaccuracy of the test, as reflected in the raw ME scores, and by the score conversion rule.

If all of these points can be dealt with satisfactorily, we may safely conclude that ME does indeed present a more reliable testing tool than MC.

6.1 Decision accuracy

The intrinsic measurement (in-)accuracy of the test methods can be estimated on the basis of score variance, as a function of the true personal probability p. There are several parameters, variables and scoring methods, however. In the interest of clarity we will restrict our analysis to the following selection:

- **k** = 2 (two-option items, e.g. True/False);
- perfect realism for ME, so that $r = p$ (the optimal test-taking strategy);
- two scoring rules: MC with correction for guessing, and ME with the score-conversion rule described earlier. For both of these rules the cutting score (5.5) corresponds to $p = 0.75$.

Furthermore, we will assume that there are 36 items in the test (for test scores from 1 to 10) and that *all* items have the same probability p. The results of this evaluation will therefore reflect the measurement accuracy as a function of p.

The variance is zero for $p = 1$ (perfect knowledge) and for $p = 0$ (complete fallacy, all items answered incorrectly), for all scoring systems. For $p = 0.5$ (zero knowledge) the variance for MC reaches a maximum, whereas the variance for ME is again zero: assuming perfect realism, all responses will be 1/**k** and all item scores will be zero. The variance for Multiple Evaluation therefore reaches a maximum somewhere between $p = 0.5$ and $p = 1.0$ as illustrated in Figure 21.

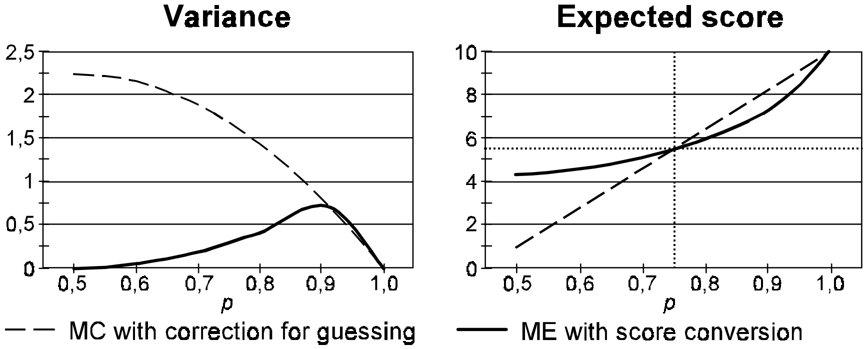


*Figure 21. Variance and expected score for **k** = 2 and p = 0.5 .. 1.0.*

For low values of p, corresponding to extreme uncertainty on the part of the student, ME is obviously a major improvement over MC. Closer to the cutting score, however, the results are less definite. The greatly reduced variance for ME must be related to the shallower angle caused by the score conversion rule: a small variance in expected score may correspond to a major difference in p.

As will be discussed in the next section, the maximum value of the variance and the corresponding value of p depend on the score conversion rule and the tolerance parameter **t**. The results illustrated in Figure 21 are valid for our original conversion rule and value for **t**.

The decision accuracy, i.e. the reliability of our ‘pass/fail’ decisions, can be estimated on the basis of two calculations at the 5% significance level. On the one hand we are interested in the probability of an ‘undeserved pass’: what low value of p corresponds to a 5% probability of a final score above our cutting score; and, on the other hand, what high value of p leaves a 5% probability of a final score below the cutting score, an ‘undeserved fail’. Although our students may not agree, we are more concerned about an ‘undeserved pass’: students just above the borderline who receive an ‘undeserved fail’ will benefit from further study for a re-sit.

Based on binomial chance distributions for p = 0.69, 0.75 and 0.87, Figure 22 shows the corresponding score distributions for MC (with correction for guessing) and ME (using our original tolerance parameter and score conversion rule). As before, we are assuming perfect realism for ME: the response $r = p$.

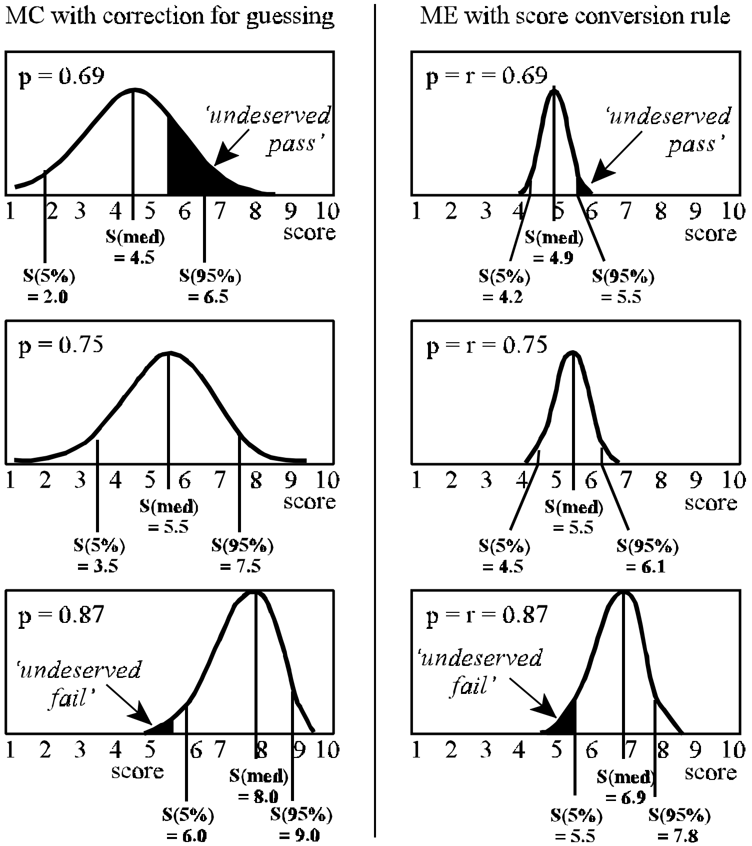


Figure 22. Score distributions for MC and ME, for three values of p.

The superiority of ME over MC is apparent. For $p = r = 0.75$ the spread of the MC scores is approximately twice that of the ME grades. More importantly for decision accuracy: for $p = 0.69$ there is 5% probability of an ME grade above our cutting score, whereas for MC 17% of the students will receive an 'undeserved pass' and the 5% score is 6.5. Or, looking at it the other way: for MC 5% of the students with p = 0.63 will receive an 'undeserved pass'.
It should be noted that this ME decision accuracy is not influenced by the score conversion rule, provided the 'cutting' value for p remains unchanged. Exactly one 'raw' ME score corresponds to this value for p, and that particular score will be converted to the desired final cutting score. The probability of receiving a higher score remains unchanged at this point. In general, the same is not true for other (secondary) cutting scores: different conversion rules will lead to different corresponding values of p. As a result, the 'measurement accuracy' and corresponding decision accuracy for other scores will depend to some extent on the score conversion rule.

At the other end of the scale, the percentage of students receiving an 'undeserved fail' is influenced to a far greater extent by the tolerance parameter **t**. This lower tail of the binomial distribution corresponds to the probability of getting a large number of items wrong, which results in a large negative score contribution. The smaller the value of **t**, the more negative these 'penalty' scores will be and the lower the final 'raw' ME score as a result. This, in turn, leads to a greater probability of an 'undeserved fail'.
With our initial choice of tolerance parameter, the results are almost identical for MC with correction for guessing and ME: the 5% 'undeserved fail' borderline for MC is $p = 0.86$ and for ME it is $p = 0.87$.

On the basis of the above we can conclude that the 'pass/fail' decision accuracy of ME is vastly superior to that of MC with respect to 'undeserved pass'. At the other end of the scale, for 'undeserved fail', we must be more careful: using our original value for the tolerance parameter, the results for MC and ME are almost identical in this aspect.

6.2 Score conversion rules and measurement accuracy

The logarithmic scoring rule required for ME results in ‘raw’ test scores that can range from extremely negative values to smaller positive values: from −40 to +10 in our case. To comply with our standard practice, these results must be converted to the range from +1 to +10. The desired cutting grade is 5.5, and we wished this to correspond to a raw ME score of 2.5. The test score conversion rule described earlier (the dotted curve in Figure 23) was based on these requirements. The test results given to the students were calculated according to this rule, so it is likely that it will have influenced their response strategy.

However, as mentioned earlier, at the low end of the scale this conversion rule is too tolerant. Responding to all items with $r = 1/\mathbf{k}$, i.e. ‘zero knowledge’, leads to a ‘raw’ score of zero, and this converts to a final grade of 4.4. Students with final grades just below our pass/fail cutting grade (5.5) may be misled into believing that they already have ‘almost sufficient’ knowledge and proficiency.

A more suitable conversion would follow an S-shaped curve: concave up to the cutting point and convex above it, as illustrated in Figure 23. At the low end of the scale (below the cutting grade) a grade of 1 or 2 indicates ‘serious fallacies’.

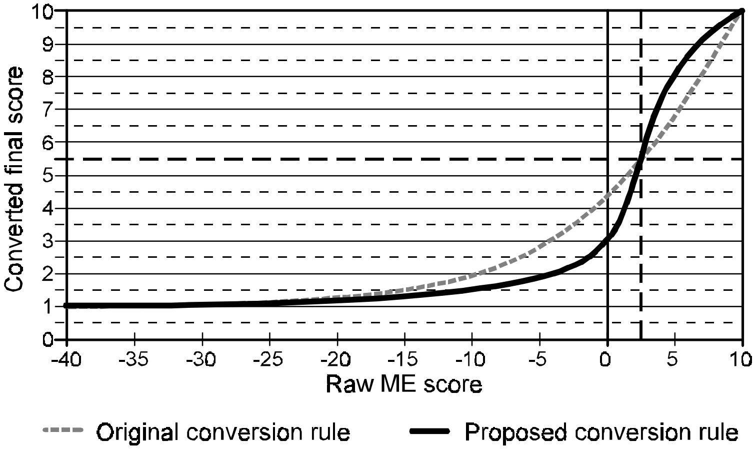


Figure 23. Desired S-shaped score conversion rule compared to the original rule (dotted line).

Such a modified score conversion rule will not impact on the pass/fail decision accuracy, as described in the previous section, since the cutting score and the corresponding value for *p* remain unchanged.

However, it *will* increase the error of measurement as reflected in the numerical value of the grades. The steeper conversion angle close to the cutting point implies that a smaller change in the raw ME score will convert into a larger change in the final grade. This leads to a larger grade variance, as illustrated in Figure 24.

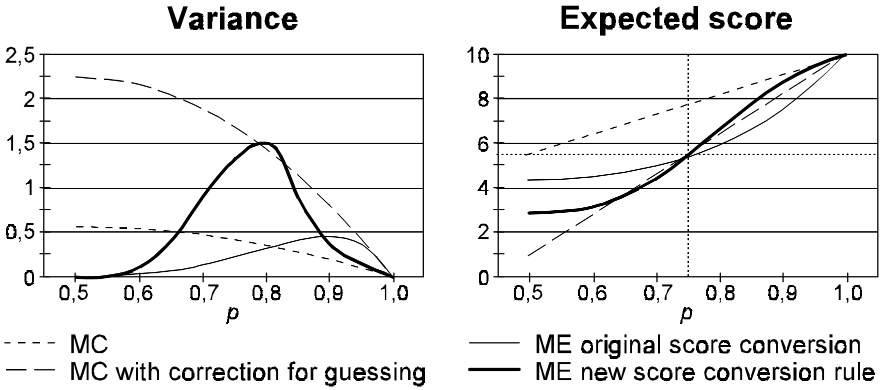


*Figure 24. Variance and expected score for **k** = 2 and p = 0.5 .. 1.0, for four scoring rules.*

Figure 25 compares the decision and measurement accuracies of this S-shaped conversion rule to those for MC with correction for guessing and those for the original conversion rule. It must be stressed that these plots were calculated for a hypothetical situation in which *all* items have the same subjective probability *p*. For the new conversion rule this gives a worst-case estimate of the measurement error near the cutting score, since the variance of the raw ME score reaches its maximum at or just above that point. In an actual test situation the subjective probabilities for the various items will be distributed over a wider range. This will lead to a smaller variance, overall, with a corresponding reduction in measurement error. In practice we may expect the rounded final grades at the cutting point to be either 5 or 6.

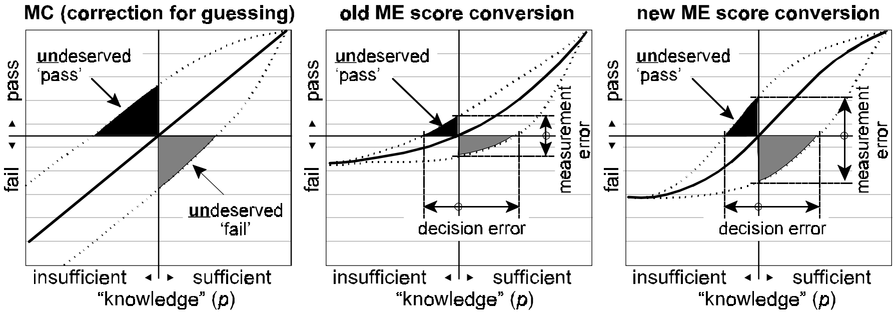


Figure 25. The dotted lines indicate the 5% probability limits around the expected final grade as a function of p. The improved decision accuracy of ME is not affected by the conversion rule.

As a final point, referring back to Figure 24, it is interesting to note the effect of this S-shaped conversion rule on the expected final grades. For low values of p, i.e. insufficient knowledge of the subject matter, the new expected final grades closely approximate the scores obtained for MC with correction for guessing - albeit with vastly improved measurement accuracy. For higher values of p, well above the cutting score, the new grades approach the results obtained for MC *without* correction for guessing.

In a sense, therefore, we may expect to get the best of both worlds. The final grade for students with very little knowledge, who might be expected to gamble on a traditional MC test, are biassed up when no correction for guessing is applied. The *expected* scores for both MC with correction for guessing and ME with the proposed score conversion rule will be more accurate in this range, but the *actual* MC scores will have a much greater variance (as shown in Figure 24). In fact the ME final grades may be expected to be even lower as p approaches $1/\mathbf{k}$, since any (overconfident) attempt to guess will decrease the final score. The expected scores plotted in Figure 24 correspond to perfect realism with $r = p$.

On the other hand, the final score for students with good knowledge are biassed downwards when correction for guessing is applied. Both MC without correction for guessing and ME with the proposed score conversion rule will be more accurate in this range.

6.3 Implementation of S-shaped score conversion rule

A simple way to implement an S-shaped curve is to actually use *two* curves, one of which is defined below the cutting point and the other above that point. Each of these curves can then be defined by three points (i.e. the two extremes and a suitably chosen intermediate point), to calculate the final grade (S) as a function of the raw test score (ME).

In our case the two extremes are (ME, S) = (–40, 1) and (10, 10). For the ME cutting score the experimental results tend to confirm our initial choice: (ME, S) = (2.5, 5.5). The first setpoint that we wish to move is the ‘zero-knowledge’ point ME = 0. The original conversion rule set this at (0, 4.38) whereas random guesswork for **k**=4 multiple-choice would give a final score of 3.25 (averaged over a sufficiently large number of items); in this range we may consider a certain amount of ‘penalty for overestimating’, and opt for a lower setpoint: (0, 3). Finally, we must select an intermediate point for the higher range; after some experimentation we settled, provisionally, on (5, 8).

On the basis of the above, we can calculate the final grade (S) as a function of the raw test score (ME) according to the following two curves:
Curve A: through the points (ME, S) = (x, y) = (–40, 1), (0, 3) and (2.5, 5.5), valid for all results up to the cutting score (2.5, 5.5); and
Curve B: through the points (ME, S) = (x, y) = (2.5, 5.5), (5, 8) and (10, 10), valid for all results above the cutting score.

Admittedly, this conversion will not be ‘smooth’ (i.e. differentiable) at the cutting point, but that is not an essential requirement. Monotonicity *is* essential, and that is guaranteed by a suitable choice of conversion functions. Figure 23 already showed this conversion rule, using the two curves defined above. For comparison, the original conversion rule is shown as a dotted line.

The two curves (below and above the cutting point) which together determine the new score conversion rule are both defined by the same basic function. Denoting the three points on each curve as (x_0, y_0), (x_1, y_1) and (x_2, y_2) gives:

$$S = K_0 + \frac{K_1}{K_2 + (x_2 - ME)} \quad \textbf{where} \tag{6.21a}$$

$$K_2 = \frac{(x_2 - x_0)(x_2 - x_1)(y_1 - y_0)}{y_0(x_2 - x_1) - y_1(x_2 - x_0) + y_2(x_1 - x_0)} \tag{6.21b}$$

$$K_1 = \frac{(y_2 - y_0)(K_2 + x_2 - x_0)K_2}{x_2 - x_0} \tag{6.21c}$$

$$K_0 = y_2 - \frac{K_1}{K_2} \tag{6.21d}$$

With the setpoints chosen above, the two sections of the conversion curve are approximately defined by:

Curve A ($-40 \leq ME \leq 2.5$): $S = 0.76 + 10.6 / (4.74 - ME)$ (6.21e)

Curve B ($2.5 \leq ME \leq 10$): $S = 13 - 37.5 / (2.5 + ME)$ (6.21f)

6.4 Old and new conversion rules compared

The difference in practice between the original and the new score conversion rules becomes apparent when we compare the two converted ME grades to the (estimated) MC scores for all tests. The results of these comparisons are given in Figure 26. Only a limited number of discrete values are possible for the ‘hard’ MC results, due to the small number of items, so these scores show as broken horizontal lines in the scattergram.

As intended, the new conversion rule cuts through the mid-range scores at a steeper angle, tending towards lower grades below the cutting score and towards higher grades above it.

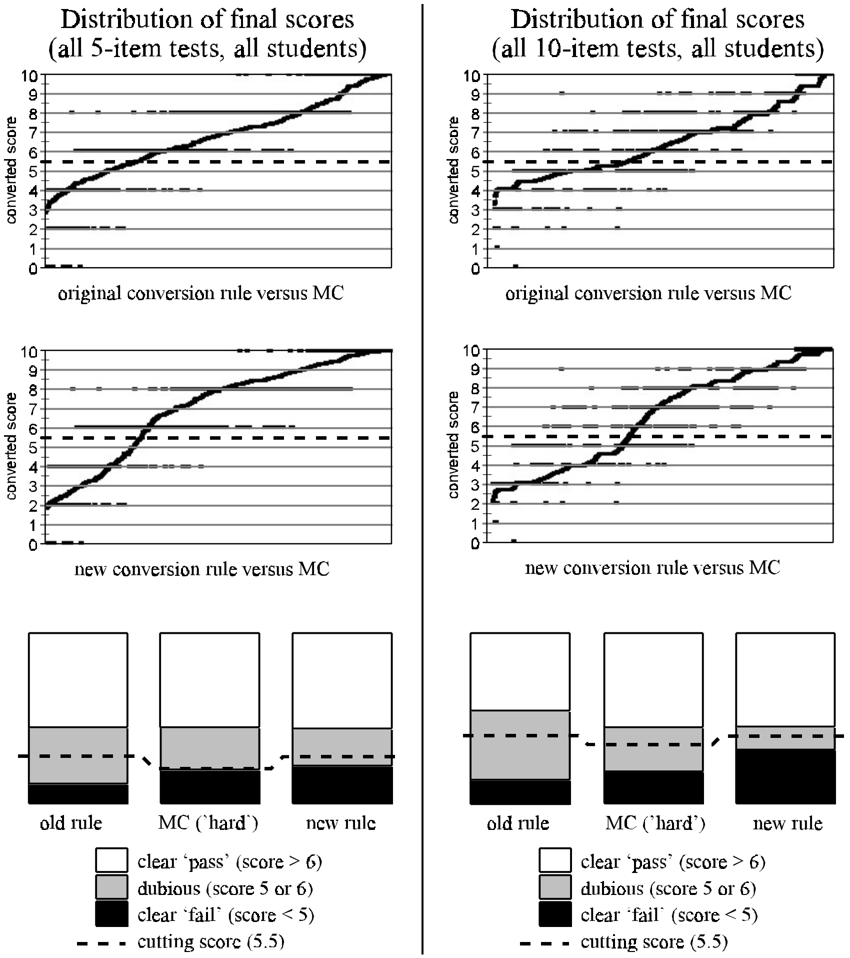


Figure 26. Comparison of scoring rules: the results for all tests and all students. In the scattergrams the 'hard' MC scores appear as broken horizontal lines. For 5-item MC, on this scale from 0 (!) to 10, only a score of 6 is considered a 'dubious pass'.

The new ME score conversion rule cuts through the dubious mid-range scores at a steeper angle than the original version, and is more 'severe' for low scores. This results in a clearer distinction between 'pass' and 'fail'.

The number of ‘dubious’ grades (5 or 6) is greatly reduced as a result. Table 13 shows the scores grouped into four ranges, from ‘clear pass’ through ‘dubious’ to ‘clear fail’. As described in the previous section, the decision accuracy for ME is sufficiently high to warrant this clearer distinction - certainly for results below the cutting score.

	computer (5 items, **k**=4)			P&P (10 items, **k**=2)		
	MC1	MEold	MEnew	MC1	MEold	MEnew
clear pass (≥7)	351	353	422	346	287	344
dubious pass (=6)	138	94	25	70	92	35
dubious fail (=5)	0	91	31	99	169	56
clear fail (≤4)	113	64	124	120	87	200

Table 13. Number of actual scores in each of four ranges, from ‘clear pass’ to ‘clear fail’, for the various score conversion rules. The new conversion rule greatly reduces the number of ‘dubious’ scores.

As in section 6.1, we can also estimate the theoretical decision accuracy on the basis of binomial chance distributions. Table 14 shows the borderline cutting response percentages, assuming perfect realism ($r = p$) for **k** = 2.

	MC1	MCcorr	MEold	MEnew
clear pass: 95% chance of score ≥7	73%	89%	93%	91%
dubious pass: 95% chance of score ≥6	63%	85%	87%	87%
nominal cutting score = 5.5	*50%*	*75%*	*78%*	*78%*
dubious fail: 95% chance of score ≤5	36%	63%	69%	69%
clear fail: 95% chance of score ≤4	27%	55%	52%	66%

Table 14. Comparison of cutting percentages for each of four ranges. MC1 is without and MCcorr is with ‘correction for guessing’.

6.5 Other sources of measurement error

Two other sources of measurement error should be considered: student attitude and the quality of the test items. Although these factors are not peculiar to ME, their effect on the final grades may be greater than for MC.

As noted earlier, students tend to be ill-calibrated initially. Good initial instruction and adequate feedback during the first few weeks are essential, but when given this assistance most students quickly learn to give realistic probability assignments. Furthermore, for certifying tests such as these the responses can be corrected for lack of realism - to some extent, at least. For the initial instruction we currently use a paper&pencil (**k** = 2) test; a translated and slightly modified version is given in Appendix 2. After a general explanation of the principles involved, as summarised in the same Appendix, students fill in this test; the results are then discussed on the basis of the notes given for each item. One lecture period (50 minutes) is sufficient for this introduction. In practice, most students do quite well on the first ‘real’ test after this instruction.

Student attitude is another matter. External factors with a detrimental effect should be avoided, and Shuford mentions the ‘cutting score’ as a possible example: students may aim for a ‘pass’ by studying only part of the subject matter in depth, and skipping the rest. This cuts both ways, however: ‘good’ students may indeed decide that enough is enough and put less effort into their study and less careful consideration into their probability assignments. On the other hand, students with an average score in the danger zone may put in more effort and more care. These effects are true of all test methods. In practice we have seen both effects occurring, but with one important difference: the ‘good’ students only attempted the minimum-effort strategy once or at most twice. They soon discovered that just one serious miscalculation on an item that they considered ‘safe’ leads to an extremely low final score. As one student told me recently: “I aim at being able to answer all the questions with at least 75% certainty, but it’s comforting to be able to answer ‘don’t know’ to questions that I *really* don’t know!”.

The other factor is the quality of the test items: the validity of any test (no matter what test method is used) will depend on this. Great care should be taken when constructing items, to ensure that they will *actually* measure what they are *intended* to measure. Given the large number of items in traditional MC tests, it is possible (and indeed even common practice) to delete 'dubious' items on the basis of low r.i.t. values.

For ME tests, however, one of the major advantages is that a considerably smaller number of items is required. This means that even greater care should be taken when constructing the items: each item is important, and deleting even one or two (on the basis of poor results) may be highly undesirable.

6.6 Future developments, avenues for further research

Our research shows that Multiple Evaluation is indeed a better method for testing than traditional Multiple Choice. The basic method for certifying tests has been developed (see Appendix 1) with provisional values for the various parameters, although some fine-tuning of these parameters must still be considered.

The method employed in this research for obtaining 'corrected-for-realism' ME scores is unsatisfactory, since it entails calculating *two* 'final' scores and then considering the higher of these two as 'valid'. It would be preferable, obviously, to have just one rule that is equally 'fair' towards both underestimators and overestimators. Our currently available data indicates that correcting the response percentages on the basis of the estimated realism score is to be preferred. Students who benefited (marginally) more from the modified tolerance parameter were all overestimators, giving a 100% response to most of the correct answers but also giving an extreme response to an *incorrect* alternative for one or two items. The marginally lower score obtained on the basis of 'corrected-for-realism' probability assignments appears more appropriate.

It should be noted that this correction is primarily intended for certifying tests, where we wish to reduce the effect on the final scores of a lack of realism. For diagnostic tests the uncorrected ME scores are probably more suitable.

From a practical point of view, the development of new software is important. Ideally, this should provide at least the following features:

For item and test construction:

- provision for importing standard file formats (e.g. Word, WordPerfect);
- user-friendly WYSIWYG editor;
- flexible options for layout (fonts, symbols, equations, graphics, tables etc.);
- compact file format, including the option ‘graphics on disk’;
- optional ‘shuffling’ of items and/or answer alternatives;
- optional setting of all relevant parameters (e.g. tolerance, score conversion);
- password protection.

For diagnostic tests:

- user-friendly (logical buttons, adequate Help file);
- suitable for Internet distribution (compact file format, system-independence);
- immediate feedback concerning subject matter (correct answer, comments);
- immediate feedback concerning performance (running ME score and realism);
- clear presentation of overall results, relevant suggestions ‘how to improve’.

For certifying tests:

- user-friendly (logical buttons, adequate Help file);
- suitable for network use (server / multi-user);
- protection for system failure (running backup on both server and user-PC);
- protection against fraud (running logfile, no feedback until results finalised);
- clear presentation of overall results *after* the test results are finalised.

For item and test evaluation (certifying tests):

- all data, for all students, password-protected in one overall file (on the server);
- graphical and numerical item analysis (presented in ‘layman’s language’);
- provision for eliminating items (e.g. on the basis of item analysis).

In an ideal world, this ‘integrated ME testing environment’ would also provide for *item banking* and *automatic test generation*.

References

[1] Abedi, J. & Bruno, J.E. (1989). Test-retest reliability of computer-based MCW-APM test scoring methods. *Journal of computer-based instruction, 16*, 29-35.

[2] Ahlgren, A. (1969). Reliability, predictive validity, and personality bias of confidence-weighted scores. (Remarks delivered in the symposium "Confidence on Achievement tests – theory, applications" at the 1969 meeting of the AERA and NCME.)
Currently available via: http://www.p-mmm.com, EXPLORERS

[3] Anderson, R.I. (1982). Computer-based confidence testing: alternatives to conventional, computer-based multiple-choice testing. *Journal of computer-based instruction, 9*, 1-9.

[4] Archer, N.S. (1962), A comparison of the conventional and two modified procedures for responding to multiple-choice items with respect to test reliability, validity, and item characteristics. *(Unpublished doctoral dissertation, Syracuse University, 1962.)*

[5] Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice*. Oxford: OUP.

[6] Baker, J.D. (1968). The Uncertain student and the understanding computer. In F. Bresson and M. de Montmollin (eds), *Programmed learning research, papers presented at the O.T.A.N. conference in Nice, May 1968,* 303-319.
Currently available via: http://www.p-mmm.com, EXPLORERS

[7] Bokhorst, F.D. (1986). Confidence-weighting and the validity of achievement tests. *Psychological Reports, 59*, 383-386.

[8] Brown, T.A. (1970). Probabilistic forecasts and reproducing scoring systems. *RM-6299-ARPA, Rand.*

Currently available via: http://www.p-mmm.com, EXPLORERS

[9] Brown, T.A. (1973). Quantifying uncertainty into numerical probabilities for the reporting of intelligence. *RM-1185-ARPA, Rand.*

Currently available via: http://www.p-mmm.com, EXPLORERS

[10] Bruno, J.E. (1986). Assessing the knowledge base of students: an information theoretic approach to testing. *Measurement and evaluation in counseling and development, 19*, 116-130.

[11] Bruno, J.E. (1993). Using testing to support feedback to support instruction: A re-examination of the role of assessment in educational organizations. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

Currently available via: http://www.p-mmm.com, EXPLORERS

[12] Coombs, C.H. (1953). On the use of objective examinations. *Educational and psychological measurement, 13*, 308-310.

[13] Coombs, C.H., Milholland, J.E. & Womer, F.B. (1956) The assessment of partial knowledge. *Educational and psychological measurement, 16*, 13-37.

[14] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré, 1937, 7*. [Translated and reprinted as "Foresight: its logical laws, its subjective sources" in H.E. Kyburg Jr. and H.E. Smokler (Eds.), *Studies in subjective probabilities.* New York: Wiley, 1964]

[15] De Finetti, B. (1962). Does it make sense to speak of good probability appraisers? In I.J. Good (Gen. Ed.), *The scientist speculates*. New York: Basic Books, 1962, 357-364.

[16] De Finetti, B. (1965). Methods of discriminating levels of partial knowledge concerning a test item. *British Journal of mathematical and statistical psychology, 18*, 87-123.

[17] De Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, 34,* 129-145.

[18] De Finetti, B. (1976). Probability: Beware of falsifications! *Scientia, 111, 1976,* 283-303.

[19] Dirkzwager, A. (1975). Computer-based testing with automatic scoring based on subjective probabilities. In Lecarme, O. & Lewis, R. (eds), *Computers in Education.* North-Holland Publishing Company, 305-311.

[20] Dirkzwager, A. (1993). A computer environment to develop valid and realistic predictions and self-assessment of knowledge with personal probabilities. In D.A. Leclercq & J.E. Bruno (eds) *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

[21] Dirkzwager, A. (1996). Testing with personal probabilities: eleven year olds can correctly estimate their personal probabilities. *Educational and psychological measurement, 56,* 957-971.
Currently available: http://www.xs4all.nl/~aried/EPMSMT.htm

[22] Dirkzwager, A. (1997). *A Bayesian Testing Paradigm: Multiple Evaluation, a feasible alternative for Multiple Choice.* (Published in [26b])
Currently available: http://www.xs4all.nl/~aried/PSYME97.htm

[23] Dirkzwager, A. (1998). *A comparison of Personal probability scoring, Certainty, Confidence scoring, Rasch estimates, Multiple Choice scoring and Realism measured with TestBet.* (Published in [26b])
Currently available: http://www.xs4all.nl/~aried/CALSART.htm

[24] Dirkzwager, A. (1999). Measurement error in measuring knowledge. (Published in [26b])

Currently available: http://www.xs4all.nl/~aried/MSMTERR.htm

[25] Dirkzwager, A. (2001). Consensus measurement in multi-participant conversations. *Kybernetes, 2001, 30, 573-587.*

Currently available: http://www.xs4all.nl/~aried/PASK97.htm

[26a] Dirkzwager, A. (2001). TestBet, Leren met Toetsen (CD-ROM), BetterSystems, Bussum 2001, ISBN 90-806315-1-5. Also available as:

[26b] Dirkzwager, A. (2001). TestBet, Learning by Testing according to the Multiple Evaluation paradigm; program, manual, founding articles (CD-ROM), ISBN 90-806315-2-3.

[27] Dirkzwager, A. (2001). Most reliable testing and assessment procedure.

Currently available: http://www.xs4all.nl/~aried

[28] Dressel, P.L. & Schmid. J. (1953). Some modifications of the multiple-choice item. *Educational and psychological measurement, 13*, 574-595.

[29] Ebel, R.L. (1965). Confidence weighting and test reliability. *Journal of educational measurement, 2*, 49-57.

[30] Ebel, R.L. (1968a). Blind guessing on objective achievement tests. *Journal of Educational Measurement, 5,* 321-325.

[31] Ebel, R.L. (1968b). Review of “Valid Confidence Testing Demonstration Kit”, *Journal of Educational Measurement, 5,* 353-354.

[32] Echternacht, G.J., Sellman, W.S., Boldt, R.F. & Young, J.D. (1971). *An evaluation of the feasibility of confidence testing as a diagnostic aid in technical training.* (Research Bulletin 71-51): Educational Testing Service.

[33] Echternacht, G.J., Boldt, R.F. & Sellmann, W.S. (1972). Personality influences on confidence test scores. *Journal of educational measurement, 9*, 235-241.

[34] Fabre, J. (1993). Subjective uncertainty and the structure of the set of all possible events. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

[35] Frary, R.B., Cross, L.H. & Lowry, S.R. (1977) Random guessing, correction for guessing and reliability of multiple-choice test scores. *Journal of experimental education, 46*, 9-15.

[36] Friedland, D.L. & Michael, W.B. (1987). The reliability of a promotional job knowledge examination scored by number of items right and by four confidence weighting procedures and its corresponding concurrent validity estimates relative to performance criterion ratings. *Educational and psychological measurement, 47*, 179-188.

[37] Gardner, W.C. (1969). Confidence testing. *USAF Instructor's Journal,* Winter 1969-1970, 4-10.

Currently available via: http://www.p-mmm.com, PIONEERS

[38] Gritten, F. & Johnson, D.M. (1941). Individual differences in judging multiple-choice questions. *Journal of educational psychology, 32*, 423-430.

[39] Hakstian, A.R. & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: testing procedures. *Journal of educational measurement, 12*, 231-239.

[40] Hambleton, R.K., Roberts, D.M. & Traub, R.E. (1970). A comparison of the reliability and validity of two methods for the assessing of partial knowledge on a multiple-choice test. *Journal of educational measurement, 7,* 75-82.

[41] Hansen, R. (1971). The influence of variables other than knowledge on probabilistic tests. *Journal of educational measurement, 8*, 9-14.

[42] Harris, D. (1969). *Testing English as a second language*. New York: McGraw-Hill.

[43] Hassmén, P. & Hunt, D.P. (1994) Human self-assessment in multiple-choice testing. *Journal of educational measurement, 31*, 149-160.

[44] Henmon, V.A.C. (1911). The relation of the time of judgment to its accuracy. *Psychological review, 18*, 186-201.

[45] Hevner, K.A. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of it. *Journal of Social Psychology, 3*, 359-362.

[46] Hopkins, K.D., Hakstian, A.R. & Hopkins, B.R. (1973). Validity and reliability consequences of confidence weighting. *Educational and psychological measurement, 33*, 135-141.

[47] Hoyt, C. (1941). Test reliability estimated by the analysis of variance. *Psychometrika, 6,* 153-160.

[48] Hughes, A. (1989). *Testing for language teachers*. Cambridge: CUP.

[49] Hunt, D.P. (1982). Effects of human self-assessment responding in learning. *Journal of applied psychology, 67*, 75-82.

[50] Hunt, D.P. (1993). Human self-assessment: theory and application to learning and testing. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

[51] Jacobs, S.S. (1971). Correlates of unwarranted confidence in responses to objective test items. *Journal of educational measurement, 8*, 15-19.

[52] Jensen, O.A. (1983). Increasing testing efficiency and effectiveness per item and per minute of testing time. *Public personnel management journal, 12*, 63-82.

[53] Klinger, A. (1997). Experimental validation of learning accomplishment. In *Proceedings of the conference on Frontiers in Education, 1997.*

Currently available via: http://www.p-mmm.com, FOUNDERS

[54] Koehler, R.A. (1974). Overconfidence on probabilistic tests. *Journal of educational measurement, 11*, 101-108.

[55] Leclercq, D.A. (1983). Confidence marking: its use in testing. *Evaluation in education, 6*, 161-287.

[56] Leclercq, D.A. (1993). Validity, reliability and acuity of self-assessment in educational testing. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

[57] Lord, F.M. (1964). The effect of random guessing on test validity. *Educational and psychological measurement, 24*, 745-747.

[58] Massengill, H.E. and Shuford, E.H., Jr. (1965). Direct vs. indirect assessment of simple knowledge structures. *ESD-TR-65-542, Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Mass.*

[59] Michael, J.J. (1968). The reliabity of multiple-choice examination under various test-taking instructions. *Journal of educational measurement, 5*, 307-314.

[60] Poizner, S., Nicewander, W.A. & Gettys, C. (1978). Alternative responses and scoring methods for multiple-choice items: An empirical study of probabalistic and ordinal response modes. *Applied psychological measurement, 2*, 83-86.

[61] Pugh, R.C. & Brunza, J.J. (1975). Effects of a confidence-weighted scoring system on measures of test reliability and validity. *Educational and psychological measurement, 35*, 73-78.

[62] Ramsey, F.P. (1926). Truth and probability. Reprinted in H.E. Kyburg Jr. and H.E. Smokler (Eds.), *Studies in subjective probabilities*. New York: Wiley, 1964.

[63] Rippey, R.M. (1970). A comparison of five different scoring functions for confidence tests. *Journal of educational measurement, 7*, 165-170.

[64] Rippey, R.M. & Voytovich, A.E. (1983) Linking knowledge, realism and diagnostic reasoning by computer-assisted confidence testing. *Journal of computer-based instruction, 9*, 88-97.

[65] Rippey, R.M. (1986). A computer program for administering and scoring confidence tests. *Behavior research methods, instruments and computers, 18*, 59-60.

[66] Roby, T.B. (1965). Belief states: a preliminary empirical study. *ESD-TDR-64-238, Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Mass.*

[67] Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association, 66,* 783-801.

[68] Shizuka, T. (2000). The validity of incorporating reading speed and response confidence in measurement of EFL reading proficiency. *Unpublished doctoral dissertation, The University of Reading.*

[69] Shuford, E.H., Jr. (1965). Cybernetic testing. *ESD-TR-65-467, Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Mass.*
Currently available via: http://www.p-mmm.com, EXPLORERS

[70] Shuford, E.H., Jr., Albert, A. & Massengill, H.E. (1966). Admissible Probability Measurement Procedures. *Psychometrika, 31*, 125-145.
Currently available via: http://www.p-mmm.com, EXPLORERS

[71] Shuford, E.H., Jr. (1967). Cybernetic testing. (Report ESD-TR-67-229, Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Mass.)
Currently available via: http://www.p-mmm.com, EXPLORERS

[72] Shuford, E.H., Jr. (1969). The logic of SCoRule testing. (Based on a paper read at the 11th annual conference of the Military testing Association, 1969).
Currently available via: http://www.p-mmm.com, EXPLORERS

[73] Shuford, E.H., Jr. & Brown, T.H. (1975). Elicitation of Personal Probabilities and their Assessment. *Instructional Science 4*, 1975, 137-188.
Currently available via: http://www.p-mmm.com, EXPLORERS

[74] Shuford, E.H., Jr. (1993). In pursuit of the fallacy: resurrecting the penalty. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.
Currently available via: http://www.p-mmm.com, EXPLORERS

[75] Shuford, E.H., Jr. (1993). *Scoring Systems, Studying and Success.* (Keynote Address at Onderwijs Research Dagen 1993, MECC Maastricht.)
Currently available via: http://p-mmm.com, EXPLORERS

[76] Sibley, W.L. (1974). An Experimental Implementation of Computer-assisted Admissible Probability Testing. *(Paper presented at 1973 Military Association Conference, San Antonio Texas, 28 October - 2 November 1973.)*

Currently available via: http://www.p-mmm.com, EXPLORERS

[77] Sieber, J.E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *Journal of personality and social psychology, 30*, 688-694.

[78] Slakter, M.J. (1968). The penalty for not guessing. *Journal of educational measurement, 5*, 141-144.

[79] Soderquist, H.O. (1936). A new method of weighting scores in a true-false test. *Journal of educational research, 30*, 290-292.

[80] Swineford (1938). The measurement of a personality trait. *Journal of Educational Psychology, 29,* 289-292.

[81] Swineford (1941). Analysis of a personality trait. *Journal of Educational Psychology, 32,* 438-444.

[82] Tarone, E. & Yule, G. (1989). *Focus on the language learner*. New York: OUP.

[83] Toda, M. (1963). Measurement of subjective probability distributions. *ESD-TDR-63-407, Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Mass.*

Currently available via: http://www.p-mmm.com, EXPLORERS

[84] Trow, W.C. (1923). The psychology of confidence. *Archives of psychology, 65*, 1-47.

[85] Van Lenthe, J. (1993). The development and evaluation of ELI, an interactive elicitation technique for subjective probability distributions. In D.A. Leclercq & J.E. Bruno (eds), *Item Banking: Interactive testing and self-assessment.* Berlin: Springer Verlag.

[86] Van Naerssen, R.F. (1961). A scale for the measurement of subjective probability. *Acta Psychologica, 20*, 159-166.

[87] Verstralen, H.H.F.M. and Verhelst, N.D. (2000). IRT models for subjective weights of options of multiple-choice questions. *CITO Measurement and Research Department Reports 2000-3.*

[88] Wiley, L.N. & Trimble, O.C. (1936). The ordinary objective test as a possible criterion of certain personality traits. *School and society, 43*, 446-448.

[89] Ziller, R. (1957). A measure of the gambling response set in objective tests. *Psychometrika, 22*, 289-292.

Appendix 1 : Suggested practical scoring formulae

Practical applications of Multiple Evaluation require the setting of the values of several parameters and the choice of a suitable score conversion rule. As a starting point for those who wish to put this method into practice, our own current choices are summarised below.

The basic logarithmic scoring rule, with a tolerance parameter, was given in Chapter 3 (3.3) :

$$s(r_c(i)) = \frac{\mathbf{ln}[(1-\mathbf{t}.\mathbf{k})\times r_c(i) + \mathbf{t}] + \mathbf{ln}(\mathbf{k})}{\mathbf{ln}(1-\mathbf{t}.\mathbf{k}+\mathbf{t}) + \mathbf{ln}(\mathbf{k})}$$

As explained in Chapter 4 the tolerance parameter, **t**, can be used to set the tolerance ratio **T**: the ratio between the maximum penalty points and the maximum score per item. The numerical values of **t** for several practical values of **T** and **k** were given in Table 3 :

	k=2	k=3	k=4	k=5
T=1	0.4999	0.1667	0.0833	0.0500
T=2	0.1910	0.0447	0.0174	0.0086
T=3	0.0804	0.0134	0.0041	0.0016
T=4	0.0362	0.0043	0.0010	0.0003

Initially we used **T** = 4 for both the **k** = 2 and **k** = 4 tests. On the basis of experience, however, we are currently using **T** = 3 for the **k** = 2 tests. Students appear distinctly more inclined to choose end-of-range percentages (0% and 100%) on these tests, whereas for **k** = 4 they usually leave a small residual percentage on all distractors.

(It is interesting to note that Shuford’s truncated logarithmic scoring rule works the other way: for **k** = 2, 3, 4 and 5 the values for **T** are approximately 5½, 3¼, 2¼ and 1¾ respectively.)

With these choices, the scoring rule can be written in numerical form.
To a good approximation,

$$\text{for } \mathbf{k} = 2: \quad s(r_c(i)) = \frac{\ln(0.839 \times r_c(i) + 0.080) + 0.693}{0.609}$$

$$\text{for } \mathbf{k} = 4: \quad s(r_c(i)) = \frac{\ln(0.996 \times r_c(i) + 0.001) + 1.386}{1.383}$$

These rules calculate ME scores on the basis of the response r, where $0 \leq r \leq 1$. When using a spreadsheet, or developing a suitable program, perfectionists can calculate more accurate numerical values on the basis of the original formula.

A measure of realism, A, is also discussed in Chapter 4.
According to Formula (4.6):

$$A = \frac{\mathbf{k} \times \sum_{i=1}^{\mathbf{m}} r_c(i) - \mathbf{m}}{\mathbf{k} \times \sum_{i=1}^{\mathbf{m}} \sum_{j=1}^{\mathbf{k}} r(i,j)^2 - \mathbf{m}}$$

where **m** is the number of items, so that the summations are over all items and all options of all items respectively. This measure can be used for feedback to the students, as described in section 5.7. Note that perfect realism corresponds to $A = 1$; $A < 1$ indicates overconfidence and $A > 1$ underconfidence. More general indications for suitable feedback are suggested in section 5.7.

Furthermore, for certifying tests this measure can be used to correct for lack of realism as discussed in section 4.4. This can be accomplished in two ways: we can modify the tolerance parameter **t**, within a limited range; or we can modify the response r, within a limited range.
In our opinion, the latter method is to be preferred, since it is effective for both under- and overestimation:

Limit A to $0.5 \leq A \leq 2$ for $\mathbf{k} = 2$, or to $0.75 \leq A \leq 2$ for $\mathbf{k} = 4$, then:
$r_{c,\text{ new}} = A \times r_c + (1-A)/\mathbf{k}$, restricted to the range $0 \leq r_{c,\text{ new}} \leq 1$.

After this ‘response correction for realism’, new ME test scores are calculated using the standard formula. These are taken as the raw ME scores, *ME* in the formulae below.

These raw scores must now be converted to final grades (S) on a standard scale, using a score conversion rule. As discussed in section 6.3, our current rule uses two curves. Each curve is defined by three points (x_0, y_0), (x_1, y_1) and (x_2, y_2) where x_n denotes the raw ME score at that point and y_n the corresponding final grade S:

$$S = K_0 + \frac{K_1}{K_2 + (x_2 - ME)} \quad \textbf{where}$$

$$K_2 = \frac{(x_2 - x_0)(x_2 - x_1)(y_1 - y_0)}{y_0(x_2 - x_1) - y_1(x_2 - x_0) + y_2(x_1 - x_0)}$$

$$K_1 = \frac{(y_2 - y_0)(K_2 + x_2 - x_0)K_2}{x_2 - x_0}$$

$$K_0 = y_2 - \frac{K_1}{K_2}$$

We are currently using the following conversion:
Curve A: through the points (ME, S) = (x, y) = (–40, 1), (0, 3) and (2.5, 5.5), valid for all results up to the cutting score (2.5, 5.5); and
Curve B: through the points (ME, S) = (x, y) = (2.5, 5.5), (5, 8) and (10, 10), valid for all results above the cutting score.

With these setpoints, the two sections are approximated by:

Curve A ($-40 \le ME \le 2.5$):	$S = 0.76 + 10.6 / (4.74 - ME)$
Curve B ($2.5 \le ME \le 10$):	$S = 13 - 37.5 / (2.5 + ME)$

When using a modified tolerance parameter to obtain a less severe maximum penalty (e.g. –30 for **k** = 2) the lower curve could be modified accordingly. The difference is minor, however, as can be seen in Figure 23 (p. 109).

Summary

The grades are calculated in three or four successive steps:

Step 1: Calculate 'realism', A, on the basis of the response r_c, where $0 \le r_c \le 1$.

$$\mathrm{A} = \frac{\mathbf{k} \times \sum_{i=1}^{\mathbf{m}} r_c(i) - \mathbf{m}}{\mathbf{k} \times \sum_{i=1}^{\mathbf{m}} \sum_{j=1}^{\mathbf{k}} r(i,j)^2 - \mathbf{m}}$$

(summing over all **m** items)

Step 2: For certifying tests, calculate 'corrected for realism' responses $r_{c,\,\mathrm{new}}$ on the basis of A. (For diagnostic testing, this step may not be desirable.)

Limit A to $0.5 \le \mathrm{A} \le 2$ for **k** = 2, or to $0.75 \le \mathrm{A} \le 2$ for **k** = 4, then:

$r_{c,\,\mathrm{new}} = \mathrm{A} \times r_c + (1-\mathrm{A})/\mathbf{k}$, restricted to the range $0 \le r_{c,\,\mathrm{new}} \le 1$.

Step 3: Calculate ME item scores on the basis of $r_{c,\,\mathrm{new}}$ (for certifying tests), or on the basis of the original responses r_c (for diagnostic testing).

$$\text{for } \mathbf{k} = 2: \quad s(r_c(i)) = \frac{\ln(0.839 \times r_c(i) + 0.080) + 0.693}{0.609}$$

$$\text{for } \mathbf{k} = 4: \quad s(r_c(i)) = \frac{\ln(0.996 \times r_c(i) + 0.001) + 1.386}{1.383}$$

Step 4: Calculate overall ME test scores as the average item score times 10, and convert these *ME* scores to final test grades *S* on a scale of 1 to 10.

Curve A ($-40 \le ME \le 2.5$): $S = 0.76 + 10.6 / (4.74 - ME)$

Curve B ($2.5 \le ME \le 10$): $S = 13 - 37.5 / (2.5 + ME)$

Appendix 2 : Initial instruction for students

For our students, one lecture period (50 minutes) is dedicated to an introduction into the principles of Multiple Evaluation. We start with a general explanation of the principles involved, on the following lines:

- Assume that we are taking a Multiple Choice test with four alternative answers for an item. Before we even look at the question there is a 25% chance of getting it right, on the basis of pure luck. Now let's say that we read it all through carefully and decide that two of the alternatives are definitely incorrect. That leaves us with two possibles, and we can't decide which is right. We must mark just one of these. If we're lucky we'll get full marks, but an unlucky guess will leave us with zero points - or maybe even a penalty.

  Surely it would be fairer if we could mark *both* of them, and then get say half of the points no matter which of the two was correct?
- Or perhaps we should go one step further. Let's simplify the situation and consider a two-alternative True/False item. Perhaps we don't consider the two alternatives equally likely: instead of 50/50 it's more like 60/40. In that case a more gradual scale might be better. For instance: full marks (10 points) if we select the correct alternative with 100% certainty; 9 points for 95%, 6 for 75%, 3 for 60% and 0 for 50% - after all, that means you don't know the answer. (They always agree, up to this point.)

  But then, of course, if you get it wrong you should get less than zero points. Let's say +3 for 60% and so –3 for 40% (they agree), +6 for 75% and –9 for 25% (surprise!), +9 for 95% and –27 for 5% (oh...!) and finally +10 for 100% and –40 for 0%.
- There is a good reason for this scale. To put it simply: you will get your highest possible score if you are realistic. When you say that you are 100% certain, you should *always* be right - so you should never get that –40 points penalty. When you say 75%, you should be right 3 out of 4 times. In that case you will score 3×6 – 9 = 9 points for those four items, which is an average of 2¼ points per item.

  However, for a 'pass' you need only average slightly more than 2½ points per item. So if you can realistically answer each and every item at over 75% you'll just scrape through. Alternatively, to go to the other extreme: if you can answer three out of ten items with 95..100% certainty and leave all others as 'don't know' (50%) you'll also just make it - *provided you get those three right!* If you get just one of them wrong you'll be in deep trouble.

– In summary: if you are (practically) certain that a statement is 'True' you should respond with 95% or 100%, and gain 9 or 10 points. Similarly, for 'certainly False' you should respond with 5% or even 0%. However: you must be almost certain, because you are risking 27 or even 40 penalty points if you are wrong!
If you think that a statement is 'True' but you feel unsure, go for 60% or 75%. Less gain (+3 or +6) but with less risk (–3 or –9).
And if you really don't know: 50%, for 0 points. "Nothing ventured, nothing lost" in this case.

– A few final remarks.
After each test you will also receive a score for 'realism'. As I said earlier, you should get three out of four right for items where you respond with 75% (or 25%, for 'False'). If you get more than that right you are underestimating your knowledge, and if you get less right you are overestimating – which will lead to a lower final score. The program will compensate for this to some extent, to protect you from your own lack of realism, but you will certainly get the highest possible score by being realistic.
Regarding the best strategy for study, note that Multiple Evaluation is not the same as Multiple Choice. For MC 'superficial learning' is the best strategy: you should learn enough about the total subject to be able to mark one option for each item with a reasonable chance of success. For ME you will be better off with a different tactic: make sure that you can answer most items with a considerable degree of certainty, and *know what you don't know* - so that you know when to respond with 50% (or maybe 40% or 60%).
Finally, regarding the best test-taking strategy: for MC gambling will usually work to your advantage so that overconfidence in your own abilities pays. For ME you should *never* 'hazard a guess': gambling is penalised, and realism is rewarded.

After this introduction the students receive written material. One page gives a brief summary of the principles outlined above, and the reverse side is a simple 10-item True/False test to give them some practical experience with the method. They are allowed 10 to 15 minutes, after which the questions and possible answers are discussed.

The items, with further comments for the final discussion, are as follows:

Are the following statements *True* or *False*? (Mark one response percentage.)

1. A traditional Multiple Choice test rewards overconfidence and gambling.

False	**0%**	**5%**	**25%**	**40%**	**50%**	**60%**	**75%**	**95%**	**100%**	**True**
if True:	*–40*	*–27*	*–9*	*–3*	*0*	*+3*	*+6*	*+9*	*+10*	
if False:	*+10*	*+9*	*+6*	*+3*	*0*	*–3*	*–9*	*–27*	*–40*	

(Comment, for the final discussion: "If you listened carefully during my introduction, you know that I consider this statement to be True. As in all tests, it is advisable to give the answer that you think the examiner wants to hear - whether you agree with him or not! If you don't agree, add your comment beside the question. I'll read that, and reply. I'll even go one step further: if I agree with your comment I'll modify your response accordingly.)

2. A test with only 10 True/False statements (50% basic probability!) can never be reliable, no matter what scoring rule is used.

   > followed by the response bar shown above <

   (Comment: this was also worked into the introduction. Yes, such a test can be sufficiently reliable, when using Multiple Evaluation.)

3. In electronics, 'digital' is the future; 'analogue' is outmoded and will linger on as a limited field for specialists.

   (Comment: note that this is an electronics course in 'Computer Systems'. I have stressed in a previous lecture that many people think this way, but that I personally do certainly not *agree. Referring back to item 1: give the answer that you think the examiner wants to hear!)*

4. The french word ‘librairie’ means ‘library’.
   (Comment: it doesn’t! It’s the french word for a book-shop. More importantly: the response percentage should reflect the student’s personal estimate of his knowledge of french. If he’s a native speaker a 0% response may be justified, but for most students only a response somewhere between a minimum of 25% and a maximum of 75% would be ‘realistic’.)

5. Yesterday evening I tossed a coin. It landed ‘heads’.
   (Comment: the only ‘correct’ response, obviously, is 50%. For a genuine test this would be a ‘nonsense question’, but it is included here as an example of an item where you really don’t know the true answer.)

6. Rounded to a whole number: $\sqrt{}1234 = 35$. **(NB: calculators not allowed!)**
   (Comment: True. However, the best response percentage depends on several factors. Have you read the question carefully? Obviously, 35^2 would give a result ending in ...5, which is not the case; but the question clearly states “rounded to a whole number”. Then, have you taken the time to think it through? If not: give a realistic answer: 50%. Finally, if you did take the time: 35^2 is too low and 36^2 is much too high, so assume that the statement is True but be careful: 75%, say?)

Items 7 to 10 are specifically tailored to our situation. They involve basic electrical theory, the number of hours students are expected to put into their study, etcetera.

After this practical demonstration and the (sometimes heated) discussion that follows, most of our students have a reasonably clear idea of what Multiple Evaluation testing entails. Some are quite realistic from the outset, and nearly all others learn to be quite realistic within one or two weeks.
In our experience, the initial 50 minutes spent on this instruction are easily recouped on the basis of reduced testing time during the rest of the course. In other follow-up courses we can then use the same system to even greater advantage, since the students are already ‘well-calibrated’.

Appendix 3 : BetterSurvey

From the outset, student evaluation of Multiple Evaluation was planned as an integral part of this research. To this end, written questionnaires were scheduled at half-term (after seven weeks) and at the end of the semester.

As a further experiment, these questionnaires were answered and scored according to a variation on Multiple Evaluation. This was based on a suggestion by Dirkzwager [25] which he called "BetVote", intended to train participants in a discussion to make good use of their estimate of the general consensus in a group. A similar method might elicit a realistic estimate of the majority opinion of a group, and so it seemed interesting to ask our (well-calibrated) students to respond to a questionnaire which was presented as a paper&pencil ME test. To be consistent with his 'TestBet' and 'BetVote', this could be termed 'BetterSurvey'.

The consecutive steps can be summarised as follows:

1. To stress the basic idea, all propositions are phrased as "Most students feel that ..." instead of the more commonly implied "I feel that...". Respondents are made clearly aware that these evaluations are intended as a measure of the general opinion of the whole group - not as a means to express personal grievances.
2. Students score each True/False item on the customary P&P k=2 scale which only allows nine discrete percentage choices: 0 - 5 - 25 - 40 - 50 - 60 - 75 - 95 - 100%.
3. The average percentage score is calculated by summing all (discrete) percentage choices and dividing by the number of respondents. This average score is taken as the initial 'key' to determine whether the statement is considered 'True' or 'False'.
4. Based on this 'key', Multiple Evaluation scores are calculated for all students (corrected for lack of realism, as in the certifying tests). This means that students who 'vote' in accordance with the group averages for all (or at least most) items will obtain higher scores; students with radically opposing overall views will have a much lower score.

5. Based on these ME scores, an individual weighting factor (W) is calculated: the individual's ME score divided by the overall average ME score for the whole group. This means that W=1 for an exactly 'average' response, less than 1 for 'dissidents' and greater than 1 for good estimators of the overall group opinion.

   In practice, we found weighting factors between 0.0 and 2.4. There were no negative ME scores, but if these had occurred they would have been taken as 0.
6. The original percentage assignments (r%) of all students for all items are 'weighted':

   $w\%(i) = (r\%(i) - 50) \times W + 50$ (%).

   This has the effect that students whose responses approximate the overall opinion of the whole group (giving them a higher ME score) will be taken 'more seriously'; radical dissidents have far less effect on the result.

   However, students who agree with the group on most counts (thereby achieving a high ME score) can throw their weight into the balance for one or two questions where they feel that their opposing view is important. In this sense, the effect closely approximates the way in which decisions are reached at a conference: people who are seen as good representatives of the overall opinion will be taken more seriously when they feel obliged to defend an opposing view. They have built up credit, which they can then use to exercise a veto-right.
7. New average scores are calculated on the basis of the weighted responses (w%). In theory this may cause the 'key' for one or more items to change from True to False or vice versa; if so, the whole process is repeated. In our case this never occurred.

In practice this system led to more clear-cut pronouncements; see Figures 17 and 18 (pp. 92 - 93).