



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Computers in Human Behavior 23 (2007) 609–627

Computers in  
Human Behavior

[www.elsevier.com/locate/comphumbeh](http://www.elsevier.com/locate/comphumbeh)

# A Bayesian sequential procedure for determining the optimal number of interrogatory examples for concept-learning

Hans J. Vos \*

*Department of Research Methodology, Faculty of Behavioral Sciences, University of Twente, Measurement and Data Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands*

Available online 8 December 2004

---

## Abstract

The purpose of this paper is to derive optimal rules for sequential decision-making in intelligent tutoring systems. In a sequential mastery test, the decision is to classify a student as a master, a nonmaster, or to continue testing and administering another item. The framework of Bayesian sequential decision theory is used; that is, optimal rules are obtained by minimizing the posterior expected losses associated with all possible decision rules at each stage of testing and using techniques of backward induction. The main advantage of this approach is that costs of testing can be taken explicitly into account. The sequential testing procedure is demonstrated for determining the optimal number of interrogatory examples for concept-learning in the Minnesota adaptive instructional system. The paper concludes with an empirical example in which, for given maximum number of interrogatory examples for concept-learning in medicine, the appropriate action is indicated at each stage of testing for different number-correct score.

© 2004 Elsevier Ltd. All rights reserved.

*Keywords:* Intelligent tutoring systems; Minnesota adaptive instructional system; Concept-learning; Sequential decision-making; Bayesian sequential principle

---

---

\* Tel.: +31 53 489 3628; fax: +31 53 489 4239.

E-mail address: [h.j.vos@utwente.nl](mailto:h.j.vos@utwente.nl).

## 1. Introduction

The term “adaptive instruction” has been in widespread use for over the past few decades (e.g., Bejar & Braun, 1994; Bork, 1984; Dijkstra, Krammer, & van Merriënboer, 1992; Gegg-Harrison, 1992; Kontos, 1985; Koper, 1995). Although different authors have defined the term in a different way, most agree that it denotes the use of strategies to adapt the appropriate amount of instruction to learning needs (Tennyson & Park, 1984). In the context of computer-based instruction, adaptive instructional programs are often qualified as intelligent tutoring systems (ITSs). Well-known examples of such systems can be found in Capell & Dannenberg (1993), De Haan & Oppenhuizen (1994), Krammer, van Merriënboer, & Maaswinkel (1994), Tennyson et al. (1984), Vasandani & Govindaraj (1994), Vos (1994), Wu & Lee (1998). Tennyson, Christensen, & Park (1984) denoted their ITS as the Minnesota Adaptive Instructional System (MAIS). The authors call their system an ITS, because it exhibits some of machine intelligence, as demonstrated by its ability to improve decision making over the history of the system as a function of accumulated information about previous students. In the literature many successful research projects have been reported on the MAIS (e.g., Park & Tennyson, 1980; Tennyson, Tennyson, & Rothen, 1980).

Initial work on the MAIS began as an attempt to design an adaptive instructional strategy for concept- or rule-learning (Tennyson, 1975). Concept-learning is the process in which subjects learn to categorize objects, processes or events; for instance, formation of diagnostic skills in medicine or psychology. According to Merrill & Tennyson (1977), concept-learning can be conceived as a two-stage process of formation of conceptual knowledge and development of procedural knowledge. An important role in concept-learning is played by expository examples (statement form), i.e., (non)examples, which organize the content in propositional format and interrogatory examples (question form), i.e., (non)examples which organize the content in interrogatory format (see Tennyson & Cocchiarella, 1986 for a complete review of the theory of concept-learning).

In the MAIS, eight basic instructional design variables directly related to specific learning processes are distinguished. In order to adapt instruction to individual learner differences (aptitudes, prior knowledge) and learning needs (amount and sequence of instruction), three of these variables are controlled by a computer-based decision strategy; namely, amount of instruction, instructional time control, and advise on learning needs (Tennyson & Christensen, 1986). The functional operation of this strategy was related to guidelines described by Novick & Lewis (1974). In the following, we shall confine ourselves only to selecting the appropriate amount of instruction in concept-learning, which can be interpreted in the MAIS as determining the optimal number of interrogatory examples for concept-learning.

Elsewhere (Vos, 1988, 1995a, 1999), the author has indicated how the problem of determining the optimal number of interrogatory examples can be reformulated as a sequential mastery problem, i.e., to declare mastery, nonmastery, or to continue testing. Sequential mastery tests are designed with the goal of maximizing the probability of making correct classification decisions (i.e., mastery and nonmastery) while at

the same time minimizing test length. For instance, Lewis & Sheehan (1990) showed that average test lengths could be reduced by half without sacrificing classification accuracy. In the MAIS, the sequential mastery problem boils down to advancing students to the next, presumably more complex concept if mastery of the present concept is clearly attained; retaining students for whom mastery of the present concept is not as clear-cut; or providing students with additional expository examples of the present concept for whom we are not sure yet whether or not they master the present concept. A new interrogatory example of the present concept is then generated for these students.

In addition, it is shown in Vos (1988, 1995a) how the computer-based decision component in the MAIS can be situated within the general framework of Bayesian decision theory (e.g., Ferguson, 1967; Lindgren, 1976), and what implicit assumptions have to be made in doing so. Also, it is shown in Vos (1988) how on some features of the MAIS decision procedure can be elaborated by using other useful results from this framework. Within a Bayesian decision-theoretic approach, optimal rules (i.e., Bayes rules) are obtained by minimizing the posterior expected losses associated with all possible decision rules. Decision rules are hereby prescriptions specifying for each possible observed response pattern what action has to be taken. The Bayes principle assumes that prior knowledge about students' true level of knowledge is available and can be characterized by a probability distribution called the prior. This prior probability represents our best prior beliefs concerning students' true level of functioning; that is, before any interrogatory example yet has been administered.

Vos (1999) also pointed out that the computer-based decision component in the MAIS can be situated within the general framework of minimax decision theory (e.g., Ferguson, 1967; Lindgren, 1976). Unlike Bayesian decision theory, specification of a prior is not necessary in applying the minimax principle. An optimal rule within the framework of minimax decision theory (i.e., a minimax rule), however, can be conceived of as a rule that is based on minimization of posterior expected loss as well (i.e., as a Bayesian rule), but under the restriction that the prior is the least favorable element of the class of priors (e.g., Ferguson, 1967). In fact, the minimax principle assumes that it is best to prepare for the worst and to establish the maximum expected loss for each possible decision rule (see van der Linden, 1981). In other words, the minimax decision rule is somewhat conservative and pessimistic (Coombs, Dawes, & Tversky, 1970). As indicated by Huynh (1980), however, the minimax principle is very attractive when the only information is the observed number-correct (i.e., number of interrogatory examples from the present concept answered correctly); that is, no group data of 'comparable' students who will take the same test is available. The minimax strategy, therefore, is sometimes also referred to as a minimum information approach (e.g., Veldhuijzen, 1982).

Furthermore, it is indicated in Vos (1995b) how the two-action classification problem in the MAIS (i.e., classifying students as either master or nonmaster) can easily be extended to the case that there are three classification decisions (classifying students as master, partial master, or nonmaster) available.

However, a major drawback of both the Bayesian and minimax framework is that it does not take costs of testing (i.e., administering another interrogatory example)

explicitly into account, which may be quite high in some real-life applications of sequential mastery testing, for instance, in developing interrogatory examples for concept-learning in medicine. The purpose of this paper, therefore, is to derive optimal rules for sequential mastery testing in the MAIS that take costs of testing explicitly into account. The framework of Bayesian sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1959) is proposed in the present paper for solving such testing problems. The main advantage of this approach is that costs of testing can be taken explicitly into account.

Although the procedures advocated in this paper are demonstrated for adapting the appropriate amount of instruction to learning needs in the MAIS, it should be emphasized that these procedures are not limited to the MAIS but, in principle, can be applied to computer-based decision components in any arbitrary ITS. In the following section, first the basic elements of the Bayesian sequential principle will be discussed and how they apply to the decision component in the MAIS. Next, it will be shown how Bayesian sequential rules can be derived. The paper concludes with demonstrating the proposed instructional strategy for computing the optimal number of interrogatory examples by an empirical example of concept-learning in medicine.

## **2. Framework of Bayesian sequential decision theory**

The following three basic elements are distinguished in the framework of Bayesian sequential decision theory. First, a loss structure evaluating the total costs and benefits for each possible combination of classification outcomes and students' (unknown) true level of functioning. Second, a probability model (i.e., measurement model) specifying the statistical relation between the observed number-correct score and students' true level of functioning. Finally, costs of administering one additional interrogatory example from the present concept must be specified.

Doing so, posterior expected losses associated with the two classification decisions (i.e., declaring mastery and nonmastery) can now be calculated at each stage of testing. The posterior expected losses are taken hereby with respect to the posterior distribution of students' true level of functioning. As far as the posterior expected loss associated with the continue testing decision concerns, this quantity is determined by averaging the posterior expected losses associated with each of the possible future decision outcomes relative to the probabilities of observing those outcomes (i.e., posterior predictive distributions). Optimal rules for the sequential mastery problem (i.e., Bayesian sequential rules) are now obtained by choosing the decision (i.e., mastery, nonmastery, or to continue testing) that minimizes posterior expected loss at each stage of testing using techniques of backward induction.

This technique starts by considering the final stage of testing (i.e., after the last interrogatory example of the present concept has been administered to the student) and then works backward to the first stage of testing (i.e., when the first interrogatory example of the present concept is going to be administered to the student). Backward induction makes use of the principle that upon breaking into an optimal

procedure at any stage, the remaining portion of the procedure is optimal when considered in its own right (DeGroot, 1970). Doing so, as pointed out by Lewis & Sheehan (1990), the action chosen at each stage of testing is optimal with respect to the entire sequential mastery testing procedure.

### 3. Notation

Within the framework of Bayesian sequential decision theory, optimal rules can be obtained without specifying a maximum number of interrogatory examples of the present concept. In the following, however, the number of interrogatory examples of the present concept is supposed to have a maximum length of  $n$  ( $n \geq 1$ ) in order to make classification decisions (i.e., mastery/nonmastery) within a reasonable period of time.

Let the observed response to an interrogatory example of the present concept at each stage of testing  $k$  ( $1 \leq k \leq n$ ) for a randomly sampled student be denoted by  $x_k$ , which can take the values 0 and 1 for, respectively, incorrect and correct responses to the  $k$ th interrogatory example.

Furthermore, let  $s_k = x_1 + \dots + x_k$  ( $0 \leq s_k \leq k$ ) denote the observed number of correct responses after  $k$  interrogatory examples of the present concept have been administered. Students' true level of functioning is unknown due to measurement and sampling error. All that is known is his/her observed number-correct score  $s_k$ . In other words, the sequential mastery test is not a perfect indicator of students' true performance. Therefore, let students' (unknown) true level of functioning be denoted by  $t$  ( $0 \leq t \leq 1$ ).

A criterion level  $t_c$  ( $0 \leq t_c \leq 1$ ), the minimum degree of students' true level of functioning required, must finally be specified in advance by the decision-maker using methods of standard-setting (e.g., Angoff, 1971; Nedelsky, 1957). A student is considered a true nonmaster and true master if his/her true level of functioning  $t$  is smaller or larger than  $t_c$ , respectively.

Assuming an observed response pattern  $(x_1, \dots, x_k)$ , two of the three basic elements of Bayesian sequential decision-making discussed earlier can now be formulated as follows: a measurement model in the form of a probability distribution,  $\text{Prob}(s_k|t)$ , relating observed number-correct score  $s_k$  to students' true level of functioning  $t$  at each stage of testing  $k$ , and a loss function describing the loss  $L(m, t)$  or  $L(n, t)$  incurred when mastery or nonmastery is declared for given  $t$ , respectively.

### 4. Threshold loss and costs of testing

Generally speaking, as noted before, a loss function evaluates the total costs and benefits of all possible classification outcomes for a student whose true level of functioning is  $t$ . These costs may concern all relevant psychological, social, and economic consequences which the decision brings along. An example of economic consequences is extra computer time associated with presenting additional expository examples. The Bayesian approach allows the decision-maker to incorporate into

the decision process the costs of misclassifications (i.e., students for whom the wrong decision is made). As in Lewis & Sheehan (1990), here the well-known threshold loss function is adopted as the loss structure involved. The choice of this loss function implies that the “seriousness” of all possible consequences of the classification decisions can be summarized by possibly different constants, one for each of the possible classification outcomes.

For the sequential mastery problem, a threshold loss function can be formulated as a natural extension of the one for the fixed-length mastery problem at each stage of testing  $k$  ( $1 \leq k \leq n$ ) as follows (see also Lewis & Sheehan, 1990).

The value  $e$  represents the costs of administering one interrogatory example of the present concept. For the sake of simplicity, following Lewis & Sheehan (1990), these costs are assumed to be equal for each classification outcome as well as for each testing occasion. Of course, these two assumptions can be relaxed in specific sequential mastery testing applications. It has been shown in the literature (e.g., Luce & Raiffa, 1957) that positive linear transformations are allowed to loss functions, that is, multiplying with a positive constant as well as adding a constant to the loss function are admissible transformations. Applying such an admissible positive linear transformation, and assuming the losses  $l_{00}$  and  $l_{11}$  associated with the correct classification outcomes are equal and take the smallest values, the threshold loss function in Table 1 was rescaled in such a way that  $l_{00}$  and  $l_{11}$  were equal to zero. That is, by subtracting  $l_{00} = l_{11}$  from each loss  $l_{ij}$  ( $i, j = 0, 1$ ). Doing so, the losses  $l_{01}$  and  $l_{10}$  associated with the incorrect classification outcomes must take positive values.

Note that no losses need to be specified in Table 1 for the continue testing option. This is because the posterior expected loss associated with the continue testing option is computed at each stage of testing as a weighted average of the posterior expected losses associated with the classification decisions (i.e., mastery/nonmastery) of future administrations of interrogatory examples with weights equal to the probabilities of observing those outcomes.

The ratio  $l_{10}/l_{01}$  is denoted as the loss ratio  $R$ , and refers to the relative losses for declaring mastery to a student whose true level of functioning is below  $t_c$  (i.e., false positive) and declaring nonmastery to a student whose true level of functioning exceeds  $t_c$  (i.e., false negative).

For assessing loss functions, or the more generally applicable utility functions, most texts on (sequential) decision theory propose lottery methods (e.g., Luce & Raiffa, 1957). But, in principle, any psychological scaling method can be used. Although helpful techniques are available, this does not mean that assessment of losses is always a simple matter. In this paper, we shall consider one such lottery method that works in decision problems with a finite number of outcomes such as

Table 1  
Table for threshold loss function at stage  $k$  ( $1 \leq k \leq n$ ) of testing

| Decision             | True level of functioning |               |
|----------------------|---------------------------|---------------|
|                      | $t \leq t_c$              | $t > t_c$     |
| Declaring nonmastery | $ke$                      | $l_{01} + ke$ |
| Declaring mastery    | $l_{10} + ke$             | $ke$          |

in the sequential mastery decision problem. In this method, it is assumed that the form of the loss structure involved is a threshold function. Therefore, only the loss parameters  $l_{ij}$  ( $i, j = 0, 1$ ) have to be assessed empirically.

Generally speaking, utility theory uses the notions of desirability of outcomes to scale the consequences of each pair of action and state of nature (see also Vos, 1988). Surely, the most desirable outcomes are the true positives and true negatives. In either case, we correctly classify the student. For this reason, it is assumed that both outcomes are equally preferred. Furthermore, if a true nonmaster is advanced, (s)he will not only lose the time required to complete the next concept, but may also become frustrated and discouraged. Therefore, let us assume that misclassifying a true master is much more desirable than misclassifying a true nonmaster.

Let  $O_{ij}$  represent the set of possible outcomes associated with making classification decision  $a_i$  (i.e., declaring mastery or nonmastery) when  $\theta_j$  is the true state of nature (i.e., true master or true nonmaster), and let  $O_{\max}$  be the most preferred outcome and  $O_{\min}$  be the least preferred outcome in this set. So, in our notation, we have  $O_{00} = O_{11} = O_{\max}$  and  $O_{10} = O_{\min}$ . Furthermore, let  $u_{ij}$  describe the utility of  $O_{ij}$ . These utilities are measured numerically on an interval scale by the following device.

Let  $u_{\max}$  and  $u_{\min}$ , the utilities of the most preferred and least preferred outcome of the possible outcomes, be assigned values 1 and 0, respectively. So, in our notation, we have  $u_{00} = u_{11} = 1$  and  $u_{10} = 0$ . Furthermore, suppose the decision-maker is indifferent between  $O_{01}$  for certain and a conceptual lottery which has probability  $p$  of realizing  $O_{\max}$  and  $(1 - p)$  of realizing  $O_{\min}$ . It is assumed (Luce & Raiffa, 1957, Chapter 2), that this indifference will occur if and only if the decision-maker has the same utility for  $O_{01}$  for sure and for the specified lottery. Since  $u_{\max} = 1$  and  $u_{\min} = 0$ , it follows that the expected utility of the lottery is  $[pu_{\max} + (1 - p)u_{\min}] = p$ . Hence, the utility of  $O_{01}$  is operationally defined as  $p$ , that is,  $u_{01} = p$ . This procedure can then be followed for each  $O_{ij}$  in turn until utilities have been assigned to each of the possible outcomes. Next, for each possible outcome  $O_{ij}$ , one can always define a suitable loss by taking the difference between the utility of the most preferred outcome and  $O_{ij}$  (Novick & Lindley, 1979).

Admittedly, specifying utility functions is not an easy task, but practical experience with the above method shows it can be done (e.g., Vrijhof, Mellenbergh, & van den Brink, 1983). If, for example, correctly classifying a student gives our decision-maker 10 “utiles” more than misclassifying a true nonmaster and only 5 “utiles” more than misclassifying a true master, then  $u_{01}$  would be 0.5. This means that misclassifying a true master is half-way between misclassifying a true nonmaster and correctly classifying a student on an interval utility scale. It follows that  $R = l_{10}/l_{01} = (u_{\max} - u_{10})/(u_{\max} - u_{01}) = (1 - 0)/(1 - 0.5) = 2$ .

An obvious disadvantage of the threshold loss function is that, as can be seen from Table 1, it assumes constant loss for students to the left or to the right of  $t_c$ , no matter how large their distance from  $t_c$ . In practice, however, errors in classification are sometimes considered to be more serious, the further away a student is from the criterion level  $t_c$ . For instance, a student who is declared a nonmaster with true level of functioning just above  $t_c$  gives the same loss as a misclassified true nonmaster with true level of functioning far above  $t_c$ . It seems more realistic to suppose that for misclassified true

nonmasters the loss is a strictly increasing function of  $t$ . Moreover, the threshold loss function shows a “threshold” at the point  $t_c$ , and this discontinuity also seems unrealistic in many cases. In the neighborhood of this point, the losses for correct and incorrect decisions should change smoothly rather than abruptly (van der Linden, 1981).

To overcome these shortcomings, van der Linden & Mellenberg (1977) proposed a continuous loss function for the fixed-length mastery problem which is a linear function of students’ true level of functioning (see also van der Linden & Vos, 1996; Vos, 1988, 1995a, 1997a, 1997b). Although a linear loss function is probably more appropriate for the sequential mastery problem, following Lewis & Sheehan (1990), in the present paper a threshold loss function is adopted for reasons of simplicity and computational efficiency. Another reason for using threshold rather than linear loss is that a linear loss function may be more appropriate in the neighborhood of  $t_c$  indeed but that the further away from  $t_c$ , however, the losses can be assumed to take more and more the same constant values again.

## 5. Measurement model

As noted before, in the present paper the well-known binomial model will be adopted for the probability that after  $k$  ( $1 \leq k \leq n$ ) interrogatory examples have been administered,  $s_k$  ( $0 \leq s_k \leq k$ ) of them have been answered correctly. Its distribution at stage  $k$  of testing for given students’ true level of functioning  $t$ ,  $\text{Prob}(s_k|t)$ , can be written as follows:

$$\text{Prob}(s_k | t) = \binom{k}{s_k} t^{s_k} (1 - t)^{k-s_k}. \quad (1)$$

Observe that  $t$  denotes the binomial distribution as well as students’ true level of functioning. The binomial model assumes that the test given to each student is a random sample of items drawn from a large (real or imaginary) item pool (Wilcox, 1981). As indicated by van den Brink (1982), when tests are criterion-referenced tests by means of sampling from item domains, such as in our sequential mastery problem, the binomial model is a natural choice for estimating the distribution of students’ number-correct score  $s_k$ , given true level of functioning  $t$ . Therefore, for each student a new random sample of interrogatory examples must be drawn in practical applications of the sequential mastery problem.

## 6. Optimizing rules for the sequential mastery problem

In this section, it will be shown how optimal rules for sequential mastery testing can be derived using the framework of Bayesian sequential decision theory. Doing so, given an observed response pattern  $(x_1, \dots, x_k)$ , first the Bayesian principle will be applied to the fixed-length mastery problem by determining which of the posterior expected losses associated with the two classification decisions is the smallest. Next, applying the Bayesian principle again, optimal rules for the sequential mastery prob-

lem are derived at each stage of testing  $k$  ( $1 \leq k \leq n$ ) by comparing this quantity with the posterior expected loss associated with the option to continue testing (i.e., administering one more interrogatory example).

### 6.1. Applying the Bayesian principle to the fixed-length mastery problem

As noted before, the Bayes rule for the fixed-length mastery problem can be found at each stage of testing  $k$  ( $1 \leq k \leq n$ ) by minimizing the posterior expected losses associated with the two classification decisions of declaring mastery or nonmastery. In doing so, as noted before, the posterior expected loss is taken with respect to the posterior distribution of  $t$ . Let  $E[L(m,t)|s_k]$  and  $E[L(n,t)|s_k]$  denote the posterior expected losses associated with these two classification decisions at stage  $k$  ( $1 \leq k \leq n$ ) of testing, respectively, given number-correct score  $s_k$ . It then follows that mastery is declared at stage  $k$  of testing when number-correct score  $s_k$  is such that

$$E[L(m,t) | s_k] < E[L(n,t) | s_k] \quad (2)$$

and that nonmastery is declared otherwise. It now can easily be verified from (1) and (2), and using Table 1, that mastery is declared at stage  $k$  of testing when  $s_k$  is such that

$$(l_{10} + ke)\text{Prob}(t \leq t_c | s_k) + (ke)\text{Prob}(t > t_c | s_k) < (ke)\text{Prob}(t \leq t_c | s_k) + (l_{01} + ke)\text{Prob}(t > t_c | s_k), \quad (3)$$

and that nonmastery is declared otherwise.  $\text{Prob}(t \leq t_c | s_k)$  denotes the cumulative posterior distribution of students' true level of functioning  $t$ . That is, the probability that students' true level of functioning  $t$  is equal to or smaller than the criterion level  $t_c$ , given an observed number-correct score  $s_k$  after  $k$  interrogatory examples have been administered, which is a decreasing function in  $s_k$ . In the MAIS,  $\text{Prob}(t > t_c | s_k) = 1 - \text{Prob}(t \leq t_c | s_k)$  (i.e., the decumulative posterior distribution of students' true level of functioning  $t$ ) is denoted as the “ $\beta$  value” or “operating level” (Tennyson et al., 1984). Rearranging terms, and using the fact that  $l_{01}, l_{10} > 0$ , it can easily be verified from (3) that mastery is declared at stage  $k$  of testing when  $s_k$  is such that

$$\text{Prob}(t \leq t_c | s_k) < 1/(1 + R), \quad (4)$$

where  $R$  denotes the loss ratio (i.e.,  $R = l_{10}/l_{01}$ ). If the inequality in (4) is not satisfied, nonmastery is declared.

Assuming a beta prior for  $t$ , it follows from an application of Bayes' theorem that under the assumed binomial model from (1), the posterior distribution of  $t$  will be a member of the beta family again (the conjugacy property, see, e.g., Lehmann, 1959). In fact, if the beta function  $B(\alpha\beta)$  with parameters  $\alpha$  and  $\beta$  ( $\alpha, \beta > 0$ ) is chosen as prior distribution and students' observed number-correct score is  $s_k$  after  $k$  interrogatory examples have been administered, then the posterior distribution of  $t$  is  $B(\alpha + s_k, k - s_k + \beta)$ . Hence, assuming a beta prior for  $t$ , it follows from (4) that mastery is declared at stage  $k$  of testing when  $s_k$  is such that

$$B(\alpha + s_k, k - s_k + \beta) < 1/(1 + R) \quad (5)$$

and that nonmastery is declared otherwise.

The beta prior might be specified as either an empirical (i.e., empirical Bayes approach) or subjective prior (i.e., subjective Bayes approach). In the first approach, empirical data from other students of the group to which the individual student belongs (i.e., ‘comparable group’) are used for estimating the parameters  $\alpha$  and  $\beta$ . Empirical procedures of estimating  $\alpha$  and  $\beta$  in the form of simple moment estimators of these parameters, based upon the mean and the reliability of the observed scores from the ‘comparable’ group, can be found in Keats & Lord (1962). In the second approach, prior knowledge about  $t$  is specified by subjective assessment. A subjective beta prior will be assumed in this paper. More specifically, the uniform distribution on the standard interval  $[0,1]$  is taken as a noninformative prior; that is, the beta distribution  $B(\alpha, \beta)$  with  $\alpha = \beta = 1$ . In other words, prior true level of functioning can take on all values between 0 and 1 with equal probability. This particular prior is used for illustrative purposes in the present paper. Its flexible form nearly always makes an approximation of prior beliefs possible (Novick & Jackson, 1974).

It then follows immediately from (5) that mastery is declared at stage  $k$  of testing when  $s_k$  is such that

$$B(1 + s_k, k - s_k + 1) < 1/(1 + R) \quad (6)$$

and that nonmastery is declared otherwise. The beta distribution has been extensively tabulated (e.g., Pearson, 1930). Normal approximations are also available (Johnson & Kotz, 1970, Section 2.4.6). In general, if  $t$  has a beta distribution with parameters  $(\alpha, \beta)$ , then this distribution can be approximated by a normal distribution with mean  $\alpha/(\alpha + \beta)$  and variance  $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$ .

## 6.2. Derivation of Bayesian sequential rules

Let  $d_k(x_1, \dots, x_k)$  denote the decision rule yielding the minimum of the posterior expected losses associated with the two classification decisions at stage  $k$  ( $1 \leq k \leq n$ ) of testing. At each stage of testing  $k$ , Bayesian sequential rules can then be found by using the following backward induction computational scheme: First, the Bayesian sequential rule at the final stage of testing  $n$  is computed. Since the option to continue testing is not available at that stage of testing, it follows immediately that the Bayesian sequential rule coincides with the Bayes rule for the fixed mastery problem (i.e.,  $d_n(x_1, \dots, x_n)$ ); that is, declare mastery of the present concept if the inequality in (6) holds for  $s_k = s_n$  and  $k = n$ ; otherwise, declare nonmastery.

Subsequently, the Bayesian sequential rule at the next to last stage of testing ( $n - 1$ ) is computed by comparing the minimum of the two classification decisions, that is,  $\min\{E[L(m,t)|s_{n-1}], E[L(n,t)|s_{n-1}]\}$ , with the posterior expected loss associated with the option to continue testing. As noted before, the posterior expected loss associated with administering one more interrogatory example at stage  $(n - 1)$  of testing, given an observed response pattern  $(x_1, \dots, x_{n-1})$ , is computed by averaging the posterior expected losses associated with each of the possible future decision out-

comes at the final stage of testing  $n$  relative to the probability of observing those outcomes (i.e., backward induction).

Let  $\text{Prob}(x_n | s_{n-1})$  denote the probability of observing response  $x_n$  ( $x_n = 0$  or  $1$ ) on the final stage of testing  $n$ , given observed number-correct score  $s_{n-1}$  on the  $(n - 1)$  previous stages of testing, then, the posterior expected loss associated with administering one more interrogatory example after  $(n - 1)$  interrogatory examples have been administered,  $E[L(c,t) | s_{n-1}]$ , is computed as follows:

$$E[L(c, t) | s_{n-1}] = \sum_{x_n=0}^{x_n=1} \min\{E[L(m, t) | s_n], E[L(n, t) | s_n]\} * \text{Prob}(x_n | s_{n-1}), \quad (7)$$

where the symbol “ $c$ ” in the left-hand side of (7) stands for continuation (i.e., administering one more interrogatory example).

Note that (7) averages the posterior expected losses associated with each of the possible future decision outcomes relative to the probability of observing those outcomes. Generally,  $\text{Prob}(x_k | s_{k-1})$  is called the posterior predictive probability of observing response  $x_k$  ( $x_k = 0$  or  $1$ ) at stage  $k$  of testing, conditional on having obtained an observed number of correct responses  $s_{k-1}$  on the  $(k - 1)$  previous stages of testing. It will be indicated later on how this conditional probability can be computed.

Given an observed response pattern  $(x_1, \dots, x_{n-1})$ , the Bayesian sequential rule at stage  $(n - 1)$  of testing is now given by: administer one more interrogatory example if  $E[L(c,t) | s_{n-1}]$  is smaller than  $\min\{E[L(m,t) | s_{n-1}], E[L(n,t) | s_{n-1}]\}$ ; otherwise, classification decision  $d_{n-1}(x_1, \dots, x_{n-1})$  is taken.

To compute the posterior expected loss associated with the option to continue testing at stage  $(n - 2)$ , the so-called risk at stage  $(n - 1)$  of testing is needed. The risk at stage  $(n - 1)$  of testing,  $\text{Risk}(x_1, \dots, x_{n-1})$ , is defined as the minimum of the posterior expected losses associated with all available decisions, that is, declare mastery, declare nonmastery, or to continue testing. In other words

$$\text{Risk}(x_1, \dots, x_{n-1}) = \min\{E[L(m, t) | s_{n-1}], E[L(n, t) | s_{n-1}], E[L(c, t) | s_{n-1}]\}. \quad (8)$$

The posterior expected loss associated with administering one more interrogatory example after  $(n - 2)$  interrogatory examples have been administered with  $s_{n-2}$  of them being answered correctly,  $E[L(c,t) | s_{n-2}]$ , can then be computed by using the following recurrent relation (i.e., computing the expected risk):

$$E[L(c, t) | s_{n-2}] = \sum_{x_{n-1}=0}^{x_{n-1}=1} \text{Risk}(x_1, \dots, x_{n-1}) * \text{Prob}(x_{n-1} | s_{n-2}). \quad (9)$$

Given an observed response pattern  $(x_1, \dots, x_{n-2})$ , the Bayesian sequential rule at stage  $(n - 2)$  of testing can now be computed analogous to the computation of the Bayesian sequential rule at stage  $(n - 1)$  of testing: administer one more interrogatory example if  $E[L(c,t) | s_{n-2}]$  is smaller than  $\min\{E[L(m,t) | s_{n-2}], E[L(n,t) | s_{n-2}]\}$ ; otherwise, classification decision  $d_{n-2}(x_1, \dots, x_{n-2})$  is taken. Following the same computational backward scheme as in determining the Bayesian sequential rules

at stages  $(n - 1)$  and  $(n - 2)$ , the Bayesian sequential rules at stages  $(n - 3), \dots, 1, 0$  are computed. The Bayesian sequential rule at stage 0 denotes the decision whether or not to administer at least one interrogatory example of the present concept.

## 7. Computation of posterior predictive probabilities

For computing the posterior expected loss associated with administering one more interrogatory example after  $(k - 1)$  interrogatory examples have been administered with  $s_{k-1}$  of them being answered correctly (i.e.,  $E[L(c, t)|s_{k-1}]$ ), as can be seen from (7) and (9), the posterior predictive probability  $\text{Prob}(x_k|s_{k-1})$  is needed. To compute this probability, a prior probability distribution must be specified. As noticed before, here the uniform distribution  $B(1, 1)$  as a special case of the beta prior  $B(\alpha, \beta)$  (i.e.,  $\alpha = \beta = 1$ ), is taken as a prior probability distribution. Another interpretation of the uniform prior is that for an artificial test length of two interrogatory examples, one response is correct (i.e.,  $\alpha = 1$ ) and one response is incorrect (i.e.,  $\beta = 1$ ), i.e., assigning equal prior odds to a correct/incorrect answer.

If  $s_{k-1}$  interrogatory examples have been answered correctly after  $(k - 1)$  interrogatory examples have been administered, and assuming a uniform prior, in combination with the binomial distribution for the psychometric model, it is known (e.g., DeGroot, 1970) that the probability on a correct response to the  $k$ th ( $1 \leq k \leq n$ ) interrogatory example (i.e.,  $\text{Prob}(x_k = 1|s_{k-1})$ ) is equal to  $(1 + s_{k-1})/(k + 1)$ . Since the probabilities on a correct and incorrect response must sum to 1, it follows immediately that the probability on an incorrect response to the  $k$ th interrogatory example (i.e.,  $\text{Prob}(x_k = 0|s_{k-1})$ ) is equal to  $[1 - (1 + s_{k-1})/(k + 1)] = (k - s_{k-1})/(k + 1)$ .

## 8. An empirical example

For illustrative purposes, the sequential testing procedure is applied to an empirical example for determining the optimal number of interrogatory examples for concept-learning in medicine. In order to determine the appropriate action under the outlined sequential testing procedure, the loss parameters  $l_{ij}$  ( $i, j = 0, 1$ ), costs  $e$  of administering one interrogatory example of the present concept, and the value of the criterion level  $t_c$  must be set in advance by the decision-maker.

Using the general backward induction scheme discussed earlier, for a given maximum number  $n$  ( $n \geq 1$ ) of interrogatory examples to be administered for the present concept, a program SEQUENTIAL was developed to determine the appropriate action (i.e., nonmastery, continuation, or mastery) at stage  $k$  ( $1 \leq k \leq n$ ) of testing for different number-correct score  $s_k$ . A copy from the program SEQUENTIAL is available from the author upon request. The program is written in Borland Pascal (Windows driven) and runs on any platform.

As an example, the appropriate action is depicted in Table 2 as a closed interval for a maximum of 20 interrogatory examples (i.e.,  $n = 20$ ) for a certain medical concept. Using the Angoff method (Angoff, 1971) as standard-setting technique, the

instructors of the course considered students as true masters if they knew at least 55% of the subject matter (i.e., medicine). Therefore,  $t_c$  was fixed at 0.55.

Furthermore, on a scale in which one unit corresponded to the constant costs of administering one interrogatory example (i.e.,  $e = 1$ ),  $l_{10}$  and  $l_{01}$  were empirically assessed by the instructors of the course (using lottery methods) as 200 and 100, respectively. These numerical values reflected the assumption that the losses corresponding to taking incorrect classification decisions were rather large relative to the costs of administering one interrogatory example. Obviously, the instructors of the course considered the loss corresponding to the false mastery decision twice as large as the loss corresponding to the false nonmastery decision (i.e.,  $R = 2$ ). Hence, the right-hand side of (6) turned out to be equal to 0.33.

Table 2 has been constructed by first computing the appropriate classification decision (i.e., mastery or nonmastery) and its associated posterior expected loss at the final stage of testing (i.e., stage 20), that is,  $d_{20}(x_1, \dots, x_{20})$  and  $\min\{E[L(m,t)|s_{20}], E[L(n,t)|s_{20}]\}$ , for all possible values of  $s_{20}$  (i.e.,  $s_{20} = 0, \dots, 20$ ). More specifically, mastery of the present concept is declared for those values of  $s_{20}$  for which the inequality in (6) holds whereas nonmastery is declared otherwise. Note that it can be inferred from Table 2 that nonmastery is declared at stage 20 if a student has answered correctly 12 or less interrogatory examples, whereas mastery is declared if 13 or more interrogatory examples have been answered correctly.

Table 2  
Appropriate action calculated by stage of testing and number-correct

| Stage of testing | Appropriate action by number-correct |              |         |
|------------------|--------------------------------------|--------------|---------|
|                  | Non-mastery                          | Continuation | Mastery |
| 0                |                                      | 0            |         |
| 1                |                                      | [0,1]        |         |
| 2                | 0                                    | [1,2]        |         |
| 3                | 0                                    | [1,3]        |         |
| 4                | [0,1]                                | [2,4]        |         |
| 5                | [0,1]                                | [2,4]        | 5       |
| 6                | [0,2]                                | [3,5]        | 6       |
| 7                | [0,2]                                | [3,5]        | [6,7]   |
| 8                | [0,3]                                | [4,6]        | [7,8]   |
| 9                | [0,4]                                | [5,7]        | [8,9]   |
| 10               | [0,4]                                | [5,7]        | [8,10]  |
| 11               | [0,5]                                | [6,8]        | [9,11]  |
| 12               | [0,5]                                | [6,8]        | [9,12]  |
| 13               | [0,6]                                | [7,9]        | [10,13] |
| 14               | [0,7]                                | [8,9]        | [10,14] |
| 15               | [0,7]                                | [8,10]       | [11,15] |
| 16               | [0,8]                                | [9,10]       | [11,16] |
| 17               | [0,9]                                | [10,11]      | [12,17] |
| 18               | [0,10]                               | 11           | [12,18] |
| 19               | [0,11]                               |              | [12,19] |
| 20               | [0,12]                               |              | [13,20] |

Likewise, the appropriate classification decision and its associated posterior expected loss has been computed after 19 interrogatory examples have been administered, that is,  $d_{19}(x_1, \dots, x_{19})$  and  $\min\{E[L(m,t)|s_{19}], E[L(n,t)|s_{19}]\}$ , for  $s_{19} = 0, \dots, 19$ . Next, the posterior expected loss corresponding with administering one more interrogatory example after 19 interrogatory examples have been administered,  $E[L(c,t)|s_{19}]$ , is computed using (7) for  $s_{19} = 0, \dots, 19$ . Another interrogatory example is administered if this value is smaller than  $\min\{E[L(m,t)|s_{19}], E[L(n,t)|s_{19}]\}$ ; otherwise, classification decision  $d_{19}(x_1, \dots, x_{19})$  is taken.

For computing the appropriate action after 18 interrogatory examples have been administered, in addition to computing  $d_{18}(x_1, \dots, x_{18})$  and  $\min\{E[L(m,t)|s_{18}], E[L(n,t)|s_{18}]\}$  for  $s_{18} = 0, \dots, 18$ , the Risk( $x_1, \dots, x_{19}$ ) at stage 19 of testing is computed using (8) for  $s_{19} = 0, \dots, 19$ . The posterior expected loss corresponding to administering one more interrogatory example after 18 interrogatory examples have been administered,  $E[L(c,t)|s_{18}]$ , can then be computed as the expected risk using (9) for  $s_{18} = 0, \dots, 18$ . Another interrogatory example is now administered if this value is smaller than  $\min\{E[L(m,t)|s_{18}], E[L(n,t)|s_{18}]\}$ ; otherwise, classification decision  $d_{18}(x_1, \dots, x_{18})$  is taken. Similarly, the appropriate action is determined at stage 17 until stage 0 of testing. The Bayesian sequential rule at stage 0 denotes the decision whether or not to administer at least one interrogatory example of the present concept.

As can be seen from Table 2, regardless of the observed number-correct score  $s_k$ , at least five interrogatory examples need to be administered before mastery of the present concept can be declared. However, in principle, nonmastery of the present concept can be declared already after administering two interrogatory examples. Of course, in practice, it can be decided to start making classification decisions only after a certain number of interrogatory examples have been administered to the student.

Also, generally a rather large number of interrogatory examples have to be answered correctly before mastery of the present concept is declared. This can be accounted for the relatively large losses corresponding to false mastery decisions relative to the losses corresponding to false nonmastery decisions (i.e., 200 relative to 100). Therefore, it seems better either to declare nonmastery of the present concept or to continue testing if the number-correct score is just on the border of declaring nonmastery/mastery. In this way, relatively large posterior expected losses associated with taking false mastery decisions can be avoided for the specific loss structure and 'cost per interrogatory example' assumed in the present example.

Finally, it can be inferred from Table 2 that with increasing number of interrogatory examples being administered, the chances of being classified as a nonmaster or master increases. This result is in accordance with our expectations, since we expect that the more interrogatory examples have been administered, the more sure we become about the classification decision to be taken.

## 9. Discussion

Optimal rules for the sequential mastery problem (nonmastery, mastery, or to continue testing) were derived within the framework of Bayesian sequential deci-

sion theory and using techniques of backward induction. The main advantage of this framework was that costs of testing could be taken explicitly into account. The procedures were demonstrated by an empirical example for determining the optimal number of interrogatory examples for concept-learning in the MAIS. For a given number of maximum interrogatory examples to be administered per concept, it was shown which steps had to be taken in order to construct a table that specifies the appropriate action (i.e., advancing to the next concept, declaring nonmastery of the present concept, or providing the student with additional expository examples of the present concept and next generating a new interrogatory example) at each stage of testing for different number-correct score. The binomial probability distribution was assumed for modeling response behavior (i.e., measurement model), whereas threshold loss was adopted for the loss function involved.

It is important to notice that the Bayesian sequential strategy is especially appropriate when costs of testing can be assumed to be quite large, such as, for instance, for concept-learning in medicine. This holds especially in case testlets (i.e., blocks of parallel interrogatory examples) are considered rather than single interrogatory examples. Another useful area of application of the Bayesian sequential strategy might be psychodiagnostics. Suppose that a new treatment (e.g., cognitive-analytic therapy) must be tested on patients suffering from some mental health problem (e.g., anorexia nervosa). Each time after having exposed a patient to the new treatment, it is desired to make a decision concerning the effectiveness/ineffectiveness of the new treatment or testing another patient. In such clinical situations, costs of testing generally are quite large and the Bayesian sequential approach might be considered as an appropriate testing strategy.

In this paper, Bayesian sequential decision theory has been used as a framework for sequential mastery testing, that is, only the stopping rule (i.e., termination criterion) is adaptive but the item (i.e., interrogatory example in the MAIS) to be administered next is selected random. However, both the item selection rule and stopping rule might be adaptive. Students' ability measured on a latent continuum is estimated after each response in this case, and the next item is selected such that its difficulty matches students' last ability estimate. Hence, as opposed to sequential mastery testing within a Bayesian sequential decision-theoretic framework with the binomial distribution as measurement model, this type of sequential mastery testing assumes that items differ in difficulty.

IRT (Item Response Theory) models are used in this approach to model the probability of answering an item correctly (i.e., the measurement model) as a function of the item's difficulty and students' ability (e.g., Hambleton, Swaminathan, & Rogers, 1991). Two IRT-based strategies have been primarily used for selecting the item to be administered next (e.g., van der Linden & Hambleton, 1997). First, the item to be administered next is the one that maximizes the amount of (Fisher's) information at students' last ability estimate. In the second IRT-based approach, the Bayesian item selection strategy, the item that minimizes the posterior variance of students' last ability estimate is administered next. In this approach, a prior distribution about students' ability must be specified.

The advantages of an IRT-based sequential testing procedure relative to a Bayesian sequential decision-theoretic approach are more flexibility and one can achieve more optimality in the sense that it is designed to choose an optimal test item at a time. However, the drawbacks are more complexity and more work needed to calibration (i.e., projecting the item difficulties and students' abilities on the same scale). As stated earlier, however, the main advantage of a Bayesian decision-theoretic approach relative to an IRT-based sequential testing procedure is that costs of testing can be taken explicitly into account.

In the present paper, as noted before, prior knowledge about  $t$  (i.e., students' true level of functioning) was assumed to be represented by a beta prior,  $B(\alpha, \beta)$ , with parameters  $\alpha$  and  $\beta$ . As shown by Keats & Lord (1962), the theoretical score distribution for a binomial model combined with a beta prior is the negative hypergeometric or beta-binomial distribution. The beta-binomial distribution is quite flexible for modeling observed score distributions via the choice of  $\alpha$  and  $\beta$  parameters. Although the focus of this paper was on establishing optimal sequential rules, however, the practice of Bayesian-based sequential testing procedures requires that model-checking must be addressed. The result stated above can be used to check if the assumptions of a binomial distribution as measurement model and a beta distribution as prior hold against the data. That is, the fit of the data to the binomial model with a beta prior can be checked by comparing the theoretical beta-binomial distribution with the empirical distribution using a  $\chi^2$  test. It may be noted that the extended beta-binomial model by Carlin & Rubin (1991) might give a better fit to the observed score distribution if the sequential mastery test is in multiple-choice format. In particular, this model might be more appropriate in case the corresponding score distribution exhibits a dearth of low scores.

When we want to implement the Bayesian-based sequential testing procedures in practice, in addition to model-checking, also the issue of sensitivity to prior specifications needs to be addressed. That is, investigating to what extent the optimal sequential rules are sensitive to small changes in the specification of the  $\alpha$  and  $\beta$  parameters of the beta distribution. As an aside, it may be noted that a well-known result from Bayesian decision theory is that when the number of responses becomes large, the parameters  $\alpha$  and  $\beta$  of the prior distribution hardly do have any longer influence on the posterior distribution of  $t$  (e.g., Gelman, Carlin, Stern, & Rubin, 1995), and thus, on the optimal sequential rules. The posterior mean of  $t$  approaches in this case namely the observed level of functioning  $s_k$ .

There are two new lines of research arising from the application of the Bayesian sequential principle for determining the optimal number of interrogatory examples. The first line is research into other loss structures than the one assumed here (i.e., threshold loss). For example, the normal ogive function (Novick & Lindley, 1979) which takes loss to be a nonlinearly function of the true level of functioning, might be a realistic representation of the losses actually incurred. This loss function does not only have realistic properties but also can be combined nicely from a statistical point of view with a standard normal distribution for the measurement model.

The second line concerns research into other prior distributions. The uniform prior on the standard interval  $[0,1]$ , as a special case of the beta distribution

$B(\alpha\beta)$  with  $\alpha = \beta = 1$ , was taken in this paper for computing the posterior predictive distribution, which, in turn was needed to calculate the posterior expected loss corresponding to continue testing by administering one more interrogatory example. It should be noted, however, that other values for the parameters  $\alpha$  and  $\beta$  ( $\alpha, \beta > 0$ ) of the beta distribution (i.e., the conjugate prior of the binomial distribution) might be used in computing the posterior predictive distribution. As an aside, it may be noted that the beta prior assumed in the MAIS can also be specified from the initial period of instruction, for instance, on the basis of the first four or six interrogatory examples (Tennyson et al., 1984). Also, although the computation of the posterior predictive distribution will become somewhat more complicated, other nonconjugate priors than the beta prior might be used. In this case, given a response pattern  $(x_1, \dots, x_k)$ , the posterior distribution  $\text{Prob}(x_k | s_{k-1})$  can be calculated directly by (numerical) integration (e.g., Gelman et al., 1995).

## References

- Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education.
- Bejar, I. I., & Braun, H. I. (1994). On the synergy between assessment and instruction: early lessons from computer-based simulations. *Machine-Mediated Learning*, 4, 5–25.
- Bork, A. (1984). Producing computer-based learning materials at the Educational Technology Center. *Journal of Computer-Based Instruction*, 11, 78–81.
- Capell, P., & Dannenberg, R. B. (1993). Instructional design and intelligent tutoring: theory and the precision of design. *Journal of Artificial Intelligence in Education*, 4, 95–121.
- Carlin, J. B., & Rubin, D. B. (1991). Summarizing multiple-choice tests using three informative statistics. *Psychological Bulletin*, 110, 338–349.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: an elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall Inc.
- DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.
- De Haan, A., & Oppenhuizen, T. (1994). SPELLER: a reflexive ITS to support the learning of second language spelling. *Computers in Human Behavior*, 10, 21–31.
- Dijkstra, S., Krammer, H. P. M., & van Merriënboer, J. J. G. (Eds.). (1992). *Instructional models in computer-based learning environments*. Heidelberg: Springer.
- Ferguson, T. S. (1967). *Mathematical statistics: a decision theoretic approach*. New York: Academic Press.
- Gegg-Harrison, T. S. (1992). Adapting instruction to the student's capabilities. *Journal of Artificial Intelligence in Education*, 3, 169–183.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Huynh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika*, 45, 167–182.
- Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: continuous univariate distributions*. Boston: Houghton Mifflin.
- Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, 27, 59–72.
- Kontos, G. (1985). Instructional computing: in search of better methods for the production of CAI lessons. *Journal of Educational Technological Systems*, 13, 121–133.

- Koper, R. (1995). PROFIL: a method for the development of multimedia courseware. *British Journal of Educational Technology*, 26, 94–108.
- Krammer, H. P. M., van Merriënboer, J. J. G., & Maaswinkel, R. M. (1994). Plan-based delivery composition in intelligent tutoring systems for introducing computer programming. *Computers in Human Behavior*, 10, 139–154.
- Lehmann, E. L. (1959). *Testing statistical hypotheses* (3rd ed.). New York: Macmillan.
- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 367–386.
- Lindgren, B. W. (1976). *Statistical theory* (3rd ed.). New York: Macmillan.
- Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
- Merrill, M. D., & Tennyson, R. D. (1977). *Concept teaching: an instructional design guide*. Engelwood Cliffs, NJ: Educational Technology.
- Nedelsky, L. (1957). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3–9.
- Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.
- Novick, M. R., & Lewis, C. (1974). Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement. CSE monograph series in evaluation* (3). Los Angeles: Center for the Study of Evaluation, University of California.
- Novick, M. R., & Lindley, D. V. (1979). The use of more realistic utility functions in educational applications. *Journal of Educational Measurement*, 15, 181–191.
- Park, O., & Tennyson, R. D. (1980). Adaptive design strategies for selecting number and presentation order of examples in coordinate concept acquisition. *Journal of Educational Psychology*, 72, 362–370.
- Pearson, K. (1930). *Tables for statisticians and biometricians*. London: Cambridge University Press.
- Tennyson, R. D. (1975). Adaptive instructional model for concept acquisition. *Educational Technology*, 15, 7–15.
- Tennyson, R. D., & Christensen, D. L. (1986). MAIS: an intelligent learning system. In D. Jonassen (Ed.), *Instructional designs for microcomputer courseware*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc..
- Tennyson, R. D., Christensen, D. L., & Park, S. I. (1984). The Minnesota Adaptive Instructional System: an intelligent CBI system. *Journal of Computer-Based Instruction*, 11, 2–13.
- Tennyson, R. D., & Cocchiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research*, 56, 40–71.
- Tennyson, R. D., & Park, O. (1984). Computer-based adaptive instructional systems: a review of empirically based models. *Machine-Mediated Learning*, 1, 129–153.
- Tennyson, C. L., Tennyson, R. D., & Rothen, W. (1980). Content structure and management strategies as design variables in concept acquisition. *Journal of Educational Psychology*, 72, 499–505.
- van den Brink, W. P. (1982). Binomial test models for domain-referenced testing. *Evaluation in Education*, 5, 166–176.
- van der Linden, W. J. (1981). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement*, 4, 469–492.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.
- van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, 1, 593–599.
- van der Linden, W. J., & Vos, H. J. (1996). A compensatory approach to optimal selection with mastery scores. *Psychometrika*, 61, 155–172.
- Vasandani, V., & Govindaraj, T. (1994). Integration of interactive interfaces with intelligent tutoring systems: an implementation. *Machine-Mediated Learning*, 4, 295–333.
- Veldhuijzen, N. H. (1982). Setting cutting scores: a minimum information approach. In W. J. van der Linden (Ed.), *Aspects of criterion-referenced measurement. Evaluation in Education: An International Review Series* (Vol. 5). pp. 141–148.

- Vos, H. J. (1988). The use of decision theory in the Minnesota Adaptive Instructional System. *Journal of Computer-Based Instruction*, 15, 65–71.
- Vos, H. J. (1994). An intelligent tutoring system for classifying students into instructional treatments with mastery scores. In F. J. Maarse, A. E. Akkerman, A. N. Brand, L. J. M. Mulder, & M. J. van der Stelt (Eds.). *Computers in psychology: applications, methods, and instrumentation* (Vol. 5, pp. 8–18). Lisse, The Netherlands: Swets & Zeitlinger.
- Vos, H. J. (1995a). Applications of Bayesian decision theory to intelligent tutoring systems. *Computers in Human Behavior*, 11, 149–162.
- Vos, H. J. (1995b). Adapting the appropriate amount of instruction to learning needs in computerized adaptive instructional systems. *Machine-Mediated Learning*, 5, 255–268.
- Vos, H. J. (1997a). Simultaneous optimization of quota-restricted selection decisions with mastery scores. *British Journal of Mathematical and Statistical Psychology*, 50, 105–125.
- Vos, H. J. (1997b). A simultaneous approach to optimizing treatment assignments with mastery scores. *Multivariate Behavioral Research*, 32, 403–433.
- Vos, H. J. (1999). Contributions of minimax theory to instructional decision making in intelligent tutoring systems. *Computers in Human Behavior*, 15, 531–548.
- Vrijhof, B. J., Mellenbergh, G. J., & van den Brink, W. P. (1983). Assessing and studying utility functions in psychometric decision theory. *Applied Psychological Measurement*, 7, 341–357.
- Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics*, 6, 3–32.
- Wu, A. K. W., & Lee, M. C. (1998). Intelligent tutoring systems as design. *Computers in Human Behavior*, 14, 209–220.