Adaptive item-based learning environments based on the item response theory: possibilities and challenges

K. Wauters,*† P. Desmet*‡ & W. Van den Noortgate*†

*iTEC, Interdisciplinary Research on Technology, Education and Communication, Katholieke Universiteit Leuven, 8500 Kortrijk, Belgium
†Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, 3000 Leuven, Belgium
‡Franitalco, Research on French, Italian and Comparative Linguistics, Katholieke Universiteit Leuven, 3000 Leuven, Belgium

Abstract

The popularity of intelligent tutoring systems (ITSs) is increasing rapidly. In order to make learning environments more efficient, researchers have been exploring the possibility of an automatic adaptation of the learning environment to the learner or the context. One of the possible adaptation techniques is adaptive item sequencing by matching the difficulty of the items to the learner’s knowledge level. This is already accomplished to a certain extent in adaptive testing environments, where the test is tailored to the person’s ability level by means of the item response theory (IRT). Even though IRT has been a prevalent computerized adaptive test (CAT) approach for decades and applying IRT in item-based ITSs could lead to similar advantages as in CAT (e.g. higher motivation and more efficient learning), research on the application of IRT in such learning environments is highly restricted or absent. The purpose of this paper was to explore the feasibility of applying IRT in adaptive item-based ITSs. Therefore, we discussed the two main challenges associated with IRT application in such learning environments: the challenge of the data set and the challenge of the algorithm. We concluded that applying IRT seems to be a viable solution for adaptive item selection in item-based ITSs provided that some modifications are implemented. Further research should shed more light on the adequacy of the proposed solutions.

Keywords

adaptive item selection, e-learning, intelligent tutoring systems, IRT, item-based learning.

Introduction

Learning environments increasingly include an electronic component, offering some additional possibilities, such as making education more accessible for the large public by creating the possibility for learners to study at their own pace, anytime and anywhere. However, one shortcoming of most learning environments is that they are static, in the sense that they provide for each learner the same information in the same structure using the same interface. Research in the field of educational sciences suggests that learners differ from each other with respect to their learning characteristics, preferences, goals and knowledge level and that these factors as well as context characteristics (e.g. location, time, etc.) have an influence on the learning effectiveness (Kelly & Tangney 2006; Verdú et al. 2008). Therefore, adaptive learning systems are developed aimed at optimizing learning conditions (Bloom 1984; Wasson 1993; Lee 2001; Verdú et al. 2008).

In the context of electronic learning environments, the term ‘adaptivity’ is used to refer to the adjustment of
one or more characteristics of the learning environment in function of the learner’s needs and preferences and/or the context. More specifically, ‘adaptivity’ can relate to three dimensions: the form, the source and the medium. The form of adaptivity refers to the adaptation techniques that are implemented (Brusilovsky 1999; Brusilovsky & Peylo 2003; Paramythis & Loidl-Reisinger 2004). Forms can be summarized in three categories: adaptive form representation, adaptive content representation and adaptive curriculum sequencing. Adaptive form representation refers to the way the content is presented to the learner and includes, for instance, whether pictures and videos are added to the text and whether links are visible or highlighted. Adaptive content representation concentrates on providing the learner with intelligent help on each step in the problem-solving process based on discovered knowledge gaps of learners. Adaptive curriculum sequencing is intended to select the optimal question at any moment in order to learn certain knowledge in an efficient and effective way. The source of adaptivity refers to the features involved in the adaptivity process and can be classified into three main categories: course/item features, such as the difficulty level and topic of the course; person features, composed of the learner’s knowledge level, motivation, cognitive load, interests and preferences; context features, which contains the time when, place from which and device on which the learner works in the learning environment; or a combination of these, for instance, the sequencing of the curriculum might be based on adapting the item’s difficulty level to the learner’s knowledge level. Regarding the medium of adaptivity, a main distinction can be made between two types of adaptive learning environments: adaptive hypermedia (AH) and intelligent tutoring systems (ITs) (Brusilovsky 1999). While AH systems are composed of a large amount of learning material presented by hypertext, ITs often provide only a limited amount of learning material. Instead, ITs intend to support a learner in the process of problem solving. ITs can be further divided into task-based and item-based ITs. Task-based ITs, such as ASSISTance and assessments (ASSISTments) (Razzak et al. 2005), are composed of substantial tasks or problems that are often tackled by means of scaffolding problems, which break the main problem down into smaller, learnable chunks. The learner tries to come to a reasonable solution by applying knowledge and skills that are needed to address these problems. Item-based ITs, such as Franel (Desmet 2006) and Spanish translation of Intelligent Evaluation System using Tests for Tele-Education (SIETTE) (Conejo et al. 2004; Guzmán & Conejo 2005), are composed of simple questions that can be combined with hints and feedback. In the remainder of this paper, we will focus on ‘adaptive curriculum sequencing in item-based ITs by matching the item difficulty level to the learner’s knowledge level’.

Recently, interests in adaptive item sequencing in learning environments have grown (e.g. Pérez-Marin et al. 2006; Leung & Li 2007). Generally, it is found that excessively difficult course materials can frustrate learners, while excessively easy course materials can cause learners to lack any sense of challenge. Learners prefer learning environments where the item selection procedure is adapted to their knowledge level. Moreover, adaptive item sequencing can lead to more efficient and effective learning. Even though several methods are adopted to match the item difficulty level to the learner’s knowledge level (e.g. Bayesian Networks; Collins et al. 1996; Millán et al. 2000; Conati et al. 2002), it seems obvious to make use of the item response theory (IRT; Van der Linden & Hambleton 1997) as it is frequently used in computerized adaptive tests (CATs) for adapting the item difficulty level to the person’s knowledge level (Lord 1970; Wainer 2000). IRT models are measurement models that specify the probability of a discrete outcome, such as the correctness of a response to an item, in terms of person and item parameters. In CAT, IRT is often used to generate a ‘calibrated item bank’, to estimate and update an examinee’s ability estimate and to decide on the next item to render. CAT can potentially decrease testing time and the number of testing items and increase measurement precision and motivation (Van der Linden & Glas 2000; Wainer 2000).

Despite the fact that IRT is a well-established approach in testing environments, for instance, within CAT, to date, the implementation of IRT within learning environments is limited. This paper will elaborate on this issue. More precisely, the aim of this paper is to identify and discuss the possibilities and challenges of extrapolating the application of IRT for adaptive item sequencing for testing applications to item-based ITs, partly based on own experiences with this type of item-based ITs. Furthermore, based on techniques used for other types of learning and testing environments, we suggest possible tracks for dealing with those challenges. A third purpose of the paper is to guide future
empirical research in this field, including the empirical evaluation of these ideas proposed in this paper.

In what follows, we will first address briefly adaptive item sequencing in testing environments since it can help to understand adaptive item sequencing in learning environments. Next, we will discuss the differences between applying IRT for adaptive item sequencing in learning environments and in testing environments. In particular, we will address the possibilities and challenges of applying IRT for this purpose in item-based ITSs. Challenges regarding the data set and the algorithms will be highlighted, and some solutions for dealing with those challenges will be suggested.

Adaptive item sequencing in testing environments

As indicated earlier, the process of tailoring a test to a person can be based on IRT. IRT models are measurement models describing responses based on person and item characteristics. The Rasch model (Rasch 1960) is the simplest IRT model and expresses the probability of observing a particular response to an item as function of the item difficulty level and the person’s ability level.

A prerequisite for adaptive item sequencing, and hence, for CAT, is to have items with a known difficulty level. Therefore, an initial development of an item bank with items of which the item difficulty level is known is needed before an adaptive test can be administered. This item bank should be large enough to include at any time an item with a difficulty level within the optimal range that was not yet presented to the person. To calibrate the items of the item bank, i.e. to estimate the item parameters (e.g. difficulty level), typically, a non-adaptive test is taken by a large sample of persons. Guidelines for the number of persons needed to obtain reliable item parameter estimates vary between 200 and 1000 persons (e.g. Huang 1996; Wainer & Mislevy 2000). Following item calibration, several steps are taken within a CAT as illustrated in Fig 1. When no information is provided about the ability level of the person, the first items that are administered are often items with an average item difficulty level. After a few items are administered, the person’s ability parameter is estimated by means of his or her answer on these items. Next, the item selection algorithm chooses the item that provides the most information about the current ability estimate as the next item. Within the framework of the Rasch model, based on which item difficulties and person abilities are directly comparable, this is an item whose difficulty level is equal to the ability level of the person. Next, the person’s ability parameter is re-estimated based on all previous items. Based on the updated estimate an additional item is selected. These steps are repeated until a certain stopping criterion is met, for instance, when the ability estimate hardly changes or when a specific number of items is reached.

Adaptive item sequencing in learning environments

The research to date has tended to focus on applying IRT for adaptive item sequencing in testing environments rather than in learning environments. Furthermore, the research on IRT in learning environments is mainly concentrated on assessment, both formative and summative (Millán et al. 2003; Guzmán & Conejo 2005; Chen et al. 2006, 2007; Leung & Li 2007; Baylari & Montazer 2009; Chen & Hsu 2008). However, some learning environments already make use of IRT to adaptively sequence the learning material (Chen et al. 2005, 2006; Chen & Chung 2008; Chen & Duh 2008; Chen & Hsu 2008). These researchers have applied original and modified IRT with success in AHs. Nevertheless, our experience with item-based ITSs,
such as Franel (Desmet 2006), shows that the application of IRT in item-based ITSs has distinct possibilities and challenges. Franel is a freely accessible item-based learning environment for French-speaking persons to learn Dutch and for Dutch-speaking persons to learn French. To optimize language learning, Franel is composed of several didactical item types, for instance, multiple choice, fill in the blank and translate, and several media features, such as video and audio. Learners are also guided towards the correct response by means of hints and elaborated feedback. This elaborated feedback, such as error-specific feedback, can depend on the answer the learner has given and therefore, can identify and correct knowledge gaps. Furthermore, learners can choose to follow a guided path or to follow their own learning path. Analysis has shown that learners tend to follow their own learning path, in which they can freely select the items they want to make. The item pool contains thousands of items. The items can be chosen from a navigation menu. The appearance of the navigation menu is not adapted to the learner, which means that every learner is presented the same navigation menu with the items in the same order. The learner can choose whether he or she wants the items to be categorized within chapters/topics (e.g. hobby, work, etc.) or whether he or she wants the items to be categorized within domains (e.g. grammar, vocabulary, etc.). Besides the topic and domain, there is no further logical order in which the items are listed within the navigation menu. Once a learner has selected an item, the learner is free to answer the item, ask for a hint or the solution or go to the next item without making any attempts.

The challenge of the data set

The difference in the data-gathering procedure of learning and testing environments has implications for IRT application in learning environments. In learning environments, data are generally collected in a less structured way, resulting in a missing data problem. Moreover, analysis and interpretation can be hampered by the skipping of items.

Missing values

In item-based ITSs, learners are free to choose the exercises they want to make. This, combined with the possibly vast amount of exercises provided within a learning environment, leads to the finding that many exercises are only made by few learners. Even though IRT can deal with structural incomplete data sets, the structure and huge amount of missing values found in the tracking and logging data of learning environments can easily lead to non-converging estimations of the IRT model parameters. Hence, the challenge of the missing values is twofold: we need to obtain the structure of data needed for item calibration and we need to obtain the amount of data needed for reliable item parameter estimation.

In practical applications of IRT, the number of items that have to be calibrated is often so large that it is practically not feasible to administer the entire set of items to one sample of learners. IRT models are able to analyse incomplete data under certain circumstances (Eggen 1993). More specifically, even though learners have not answered all items, a common measurement scale can be constructed on which all items can be located. There is, however, a problem regarding the structure of the data if there is no overlap in items solved by one or several persons with those solved by other persons. Without making additional assumptions, this lack of overlap does not allow assessing the difficulty of the items and/or the ability of the persons since in IRT, this is performed relative to the other items or persons. Unfortunately, this lack in overlap might be encountered in learning environments, especially if learners are completely free in selecting items from a large item pool, as is the case in Franel.

A method that can be applied to obtain the structure needed for item calibration is to administer all learners some common items. This would result in a calibration design such as the one presented in Fig 2. This incom-
complete calibration design with a common items anchor has the advantage that neither the equivalence of learners nor the equivalence of items needs to be assumed. Additionally, an advantage for the course creator is that he or she has some control over the item administration, such that non-overlapping parts in the data set can be avoided. A drawback of the approach, however, is that at the end, we have a lot of information about the common items but much less about the other items. Moreover, while item development is a very expensive and time intensive enterprise, not all items will be used to the same degree, and possibly, some items will hardly be used. This problem of unequal use of items might also occur in learning environments in which learners can choose items freely out of a fixed navigation menu but select items according to a certain pattern (for instance, learners try to solve the items in the order they are listed within the navigation menu). A solution to this problem is to vary the order of the items as listed in the learning environment. For instance, if all learners select some items from the first topic, less items from the second topic, even less items from the third topic, etc., systematically rotating the order in which the topics are presented in the navigation menu (last topic comes first, first second, second third and so on) could lead to a calibration design similar to the one presented in Fig 3. Applying this incomplete calibration design with a block-interlaced anchor within learning environments has two main advantages: an anchor–item effect is achieved while the number of administrations across items is kept more equal and both the course creator and the learner have some control over the item administration. A drawback is that it is still possible, though less likely, to end up with a data set with non-overlapping parts.

The amount of data refers to the data-centric character of IRT models, which implies that enough data are required to estimate the item parameters. Some researchers have used prior calibration in learning environments for item difficulty estimation (Tai et al. 2001; Guzmán & Conejo 2005; Chen & Duh 2008). However, this is a time-consuming and costly procedure and therefore less appropriate for learning environments with lots of items. Alternatively, online calibration can be employed in which both the learner’s knowledge level and the difficulty level of the new items are estimated during the learning sessions. Makransky and Glas (2010) explored a continuous updating strategy that yields a low mean average error for the ability estimate. In this continuous updating strategy, items are
initially randomly selected. After each exposure, the items are (re)calibrated. Only when an item is administered a certain amount of times does it become eligible for adaptive item selection. A second issue is that more persons are required to achieve the same level of precision with larger item banks. Besides, when the number of available items becomes large, which is the case in our item-based ITS, it will take longer for randomly selected items to be administered a certain amount of times in order to become eligible for adaptive item selection. A method to influence the administration or exposure rate of items, such that the amount of data needed for reliable item calibration is obtained faster than with random item selection, is by applying an item exposure control algorithm. Item exposure control methods are often applied in CAT (Georgiadou et al. 2007). The underlying philosophy of CAT, namely selecting items that provide the most information about the current ability estimate, leads to overexposure of the most informative items, while other items, which provide less information, are rarely selected. Several item exposure control strategies have been proposed to prevent overexposure of some items and to increase the use rate of rarely or never selected items (Georgiadou et al. 2007). Therefore, in CAT, the objective of such item exposure control strategies is to equalize the frequency with which all items are administered. Yet, the objective of item exposure control strategies in item-based ITSs would rather be increasing the administration frequency of items that are close to being reliably calibrated. A major advantage of item exposure control strategies in item-based ITSs is the decrease in the time needed for items to become eligible for adaptive item selection. An item exposure control method that imposes item-ineligibility constraints on the assembly of shadow tests seems to be a reasonable solution that can be applied within item-based ITSs (Van der Linden & Veldkamp 2004, 2007). So far, this method is only implemented in testing environments. In this item exposure control strategy, the selection of each new item is preceded by an online assembly of a shadow test. This implies that items are not selected directly from the item pool but from a shadow test. The shadow test is a full-sized test that is optimal at the person’s current ability estimate and contains all items, administered or not, that meet all the constraints of the adaptive test. The optimal item at the ability estimate from the free items in the shadow test is subsequently selected for administration. After the item is administered, the items in the shadow test that are not yet administered are returned to the item pool, the ability estimate is updated and the procedure is repeated. An asset of this method is that it can deal with a large range of constraints as long as there is a computer algorithm available. Hence, an item-ineligibility constraint can be implemented that reduces the exposure of overexposed items, that has a positive effect on the exposure rate of underexposed items (Van der Linden & Chang 2003; Van der Linden & Veldkamp 2004) or that, in the light of item-based ITSs, has a positive effect on the exposure rate of items that have an exposure rate close to one required for reliable item calibration.

Skipped items
Besides the large amount and pattern of missing values, intentionally skipping items may pose problems. In Franel, learners are not constrained to answer each selected item. The decision of skipping an item may depend on the learner’s knowledge level and on the item difficulty, making the missing data not missing at random or missing completely at random. Research has indicated that treating the omitted responses as wrong is not appropriate (Lord 1983; Ludlow & O’Leary 1999; De Ayala et al. 2001). Goegebeur et al. (2006) proposed an IRT model for dealing with this kind of omitted responses, more specifically, responses missing not at random (MNAR) in a testing environment. This IRT model models missing observations simultaneously with the observed responses by threatening the data as if there were three response categories: correct, incorrect and no response. The model also accounts for test speed. Goegebeur et al. (2006) applied this model to the Chilean Sistema di Medición de la Calidad de la Educación mathematics test data set and results show that the parameter estimates are improved with this model compared with the estimates based on an extended one-parameter logistic model of complete profiles. However, the authors emphasize that caution is in order since the results can be strongly affected by the mechanism underlying the MNAR framework.

The challenge of the algorithm
Next to the challenge of the data set, the IRT application in learning environments is confronted with a second
challenge, namely the challenge of the algorithm. This challenge is partly derived from the difference in the data-gathering procedure between learning and testing environments and partly derived from the difference in the objective of learning and testing environments. We divide the challenge of the algorithm into three sub-problems, namely, the problem in estimating the item difficulty level, the problem in estimating the learner’s ability level and the problem in defining the item selection algorithm.

**Item difficulty estimation**

Because of the different problems facing IRT-based item calibration in learning environments that arise from the data-gathering procedure, more specifically, the missing values and the skipped items, other approaches to estimate the item difficulty level have been proposed.

A simple approach to estimate item difficulty that is not directly related to the IRT model is the number of learners who have answered the item correctly divided by the number of learners who have answered the item. This proportion of correct answers has the benefit that it is not based on a prior study but can be calculated online. The lower the proportion of persons who answer an item correct, the more difficult the item is. Johns et al. (2006) have compared the item difficulty levels obtained by training an IRT model with the proportion correct answers. Even though the correlation between those two measures of item difficulty across 70 items was relatively high ($r = +0.68$), the proportion of correct answers is subject to the knowledge level of the learners who have answered that item and to the amount of learners who have answered that item. Hence, although attractive due to its simplicity, the approach throws away an important strength of IRT: the proportion of correct answers for an item depends on the sample of persons who answered the item, and therefore, these proportions are only comparable over items if the group of persons is comparable. Another method that has been used to estimate item difficulty in AHs is by means of the learner’s feedback on simple questions after each item (e.g. Chen et al. 2005, 2006; Chen & Duh 2008), such as ‘Do you understand the content of the recommended course materials?’ After a learner has given feedback, scores are aggregated with those of other learners who previously answered this question. The new difficulty level of the course material is based on a weighted linear combination of the course difficulty as defined by course experts and the course difficulty determined from collaborative feedback of the learners. The difficulty parameters slowly approach a steady value as the number of learners increases. Another approach to obtain item parameter estimates is allowing subject domain experts to estimate the value of the difficulty parameter (Yao 1991; Linacre 2000; Fernandez 2003; Lu et al. 2007). There is some evidence in the measurement literature that test specialists are capable of estimating item difficulties with reasonable accuracy (e.g. Chalifour & Powers 1989), although other studies found contradictory results (Hambleton et al. 1998). Another method that has been used in testing environments to estimate the item difficulty level of new items is the paired comparison method (Ozaki & Toyoda 2006, 2009). In this method, items for which the difficulty parameter has to be estimated are compared with multiple items, of which the item difficulty parameter is known. This is performed sequentially or simultaneously, i.e. a one-to-one comparison or a one-to-many comparison, respectively. Learner feedback, expert feedback and paired comparison have two common limitations. First, these three alternative estimation approaches are less applicable for the calibration of an entire, large, already existing item bank since asking the learner for feedback after each item administration requires both considerable time and mental effort, possibly interrupting the learning process. However, these estimation approaches seem to be viable solutions for estimating the item difficulty level of new added items. Second, these estimation approaches are more subjective than IRT-based calibration.

**Ability estimation**

The objective of adaptive learning environments is to foster efficient learning by providing learners a personalized course/item sequence. Obtaining the learner’s ability estimate at an early stage of the learning process and following the learning curve adequately is required in order for adaptive learning environments to be efficient. Hence, the problem of the ability estimation is twofold. On the one hand, we need to estimate the learner’s ability level when little information is provided, which is referred to as the cold start problem. On the other hand, we need to be able to follow the progression of the learner.
The **cold start problem** refers to the situation where a learner starts working in the learning environment. After having answered a few items, the ability estimate is not very accurate, reducing the efficiency of adaptive item selection. Therefore, two partial solutions can be provided. On the one hand, it can be feasible to use personal information, such as previous education, occupation, age and gender to obtain and improve the learner’s ability estimate. This technique can be extended by assigning weights to these criteria to model their relative importance (Masthoff 2004). On the other hand, it is advisable to apply other item selection methods than the one based on maximizing the Fisher information at the current estimated ability level because using Fisher information when the estimated ability level is not close to the true ability level could be less efficient than assumed. A potential solution to this problem is modifying the item selection algorithm by taking into consideration the uncertainty of the estimated ability level. Several modified item selection algorithms have been proposed: Fisher interval information (Veerkamp & Berger 1997), Fisher information with a posterior distribution (Van der Linden 1998), Kullback–Leibler information (Chang & Ying 1996) and Kullback–Leibler information with a posterior distribution (Chang & Ying 1996). For a comparison of these item selection algorithms, see Chen et al. (2000).

The **problem of the evolving ability level** arises because learning is likely to take place on the basis of hints and feedback. Besides the change in the ability level within a learning session, the ability of the learner might increase between two successive learning sessions, for instance, due to learning in other courses they follow, or might decrease, for instance, due to forgetting. This monitoring of the learner’s progress is an important research focus in the field of educational measurement. One method is progress testing. In progress testing, tests are frequently administered to allow for a quick intervention when atypical growth patterns are observed. However, a major drawback of this approach is that it requires regular ability assessments that should be long enough to be accurate, making it less appropriate for learning environments. Another method to model the learner’s ability progress is by updating the learner’s knowledge level after each item administration as is the case in the Elo rating system (Elo 1978), which was recently implemented in the educational field (Brinkhuis & Maris 2009). In item-based ITSs, the Elo rating system can be seen as an instance of paired comparison where the learner is seen as a player and the item is seen as its opponent. In the formula of Brinkhuis and Maris (2009), the new ability rating after an item administration is function of the pre-administration rating, a weight given to the new observation and the difference between the actual score on the new observation and the expected score on the new observation. This expected score is calculated by means of the Rasch model. This means that when the difference between the expected score and the observed score is high, the change in the ability estimate is high. A merit of this algorithm is that it makes it possible to quickly follow changing knowledge levels and rapidly perceive an atypical growth pattern.

**Item selection algorithm**

The difference in objectives between a testing and a learning environment also asks for a revision of the item selection algorithm as it is implemented in testing environments.

The objective of a testing environment is to measure as precisely as possible the person’s ability level. It can be shown that for the Rasch model, items with a success probability of 50% are optimally informative. However, the administration of items for which the person has a success rate of only 50% can decrease the person’s motivation (Andrich 1995) and increase test anxiety (Rocklin & Thompson 1985), possibly resulting in an underestimation of the person’s ability (Betz & Weiss 1976; Rocklin & Thompson 1985). Administering easier items in order to increase motivation and decrease test anxiety is non-optimal from a psychometric point of view. However, the impact on the measurement precision and test length is modest (Bergstorm et al. 1992; Eggen & Verschoor 2006). Another method applied in CAT resulting in a decrease in test anxiety and an increase in motivation and performance is allowing the person to select the item difficulty level of the next item from among a number of difficulty levels (Wise et al. 1992; Vispoel & Coffman 1994; Roos et al. 1997). Although most persons choose a difficulty level that lies close to their estimated ability level (Johnson et al. 1991; Wise et al. 1992), a problem with these self-adapted tests still remains that a few persons choose a difficulty level that is not well-matched to their ability, resulting in a flaw in the psychometric demands.
In contrast, the objective of a learning environment is to optimize learning efficiency. The effect of motivation herein is even of greater importance than in testing and especially than in high-stake testing. Unmotivated learners will likely stop using the learning environment. Hence, especially for learning environments items for which the person has a success probability of above 50% should be considered. Next to increasing the success probability in order to increase motivation, it is possible to give the learner some control over the item difficulty level, comparable to self-adapted tests. Besides motivation, the learning outcome should also be kept optimal. Research results regarding the effect of item difficulty on the learning progress are not conclusive. Some studies have indicated that learners learn more from easy items (Pavlik & Anderson 2008), while some theories suggest that learners learn more from items that are slightly more difficult given the learner’s ability level (Hootsen et al. 2007), and yet, other studies have indicated that learners learn as much from easy groups of learning opportunities as from difficult ones (Feng et al. 2008).

Discussion and conclusion

Because IRT has been a prevalent approach for adaptive item sequencing in testing environments, the question that is tackled within this paper is whether IRT is suitable for adaptive item sequencing in item-based ITSs. Based on own experiences with item-based ITSs and ideas borrowed from testing environments and other types of learning environments, we found some challenges in applying IRT for adaptive item sequencing in item-based ITSs and made some suggestions to handle these challenges (see Table 1).

A first main challenge that is brought up is that the structure of the data set and the extent of missing values can make item difficulty estimation problematic. Solutions that are suggested include the implementation of a calibration design, such as a common items design and a block-interlaced anchoring design. The former has the advantages that neither the equivalence of learners nor the equivalence of items needs to be assumed. However, the common items are administered more frequently and are therefore more reliably estimated than the non-common items. The latter has the same advantages as the common items design without its disadvantage. Because learners seem to like some control over the learning system (e.g. select the chapter), a block-interlaced anchoring design seems to be preferable in item-based ITSs as it provides both the course creator and the learner some control over the item administration. Furthermore, an online calibration method to evolve from random item administration to adaptive item administration is suggested. The advantage of this continuous updating strategy is that items can easily be added in the learning environment without the requirement of prior calibration while keeping the measurement error of the learner’s knowledge level low. Finally, an item exposure control algorithm can be specified that makes items for which the item exposure rate is close to the one required for reliable item difficulty estimation, eligible for adaptive item administration. This item exposure control method that imposes item ineligibility constraints has the benefit of reducing the time needed for item calibration. Combining the proposed solutions might also be considered. For example, when we want to measure a learner’s ability level while calibrating an item bank, it might be reasonable to combine a continuous updating calibration strategy with an item exposure control algorithm. In such situations, a part of the learning session is composed of randomly selected items with an unknown difficulty level. In another part, items are selected that have been administered already a certain amount of times but from which the difficulty level cannot yet be reliably estimated. Items are selected that are closest to being reliably calibrated, and the item difficulty parameter estimate is updated after each item administration. In a last part, adaptive item selection can be applied using items for which the difficulty level is already sufficiently known. A difficulty is that we do not know in advance how many items a learner will complete in one learning session.

A second main challenge, largely intertwined with the challenge of the data set, concerns the implemented algorithms, which is composed of the item difficulty estimation, the learner’s ability level estimation and the item selection algorithm. The alternative applicable techniques for item difficulty estimation that are proposed in this paper are proportion correct, learner’s feedback, expert rating and paired comparison. Compared with IRT-based calibration, these alternative techniques are more sample-dependent (proportion correct) or more subjective because persons have to make a decision (learner’s feedback, expert rating and paired comparison). Besides, learner’s feedback, expert rating and
<table>
<thead>
<tr>
<th>Challenge</th>
<th>Problem</th>
<th>Proposed solution</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Applications</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common items design</td>
<td></td>
<td></td>
<td>* No assumption needed on the equivalence of learners and items</td>
<td>* Non-equal administration of items</td>
<td>* The calibration of an item bank</td>
<td>Eggen 1993</td>
</tr>
<tr>
<td>Block-interlaced anchoring design</td>
<td></td>
<td></td>
<td>* No assumption needed on the equivalence of learners and items</td>
<td></td>
<td></td>
<td>Eggen 1993</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* More equal administration of all items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of data</td>
<td>Continuous updating strategy</td>
<td></td>
<td>* Easiness in adding items to item pool without prior calibration</td>
<td>* Length of learning session is preferably known</td>
<td>* The calibration of an item bank</td>
<td>Makransky and Glas 2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>* Low measurement error in ability estimation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skipped items</td>
<td>Item-ineligibility constraints on the assembly of shadow tests</td>
<td></td>
<td>* Faster calibration</td>
<td></td>
<td>* The calibration of an item bank</td>
<td>Van der Linden &amp; Veldkamp 2004, 2007</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Simultaneous modeling of the missing observations with the observed responses</td>
<td></td>
<td>* More accurate parameter estimates</td>
<td>* More complex model and hence more data required for estimation</td>
<td>* Estimation of learner's ability level</td>
<td>Goegebeur et al. 2006</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Item difficulty</td>
<td>Proportion correct</td>
<td>* Always estimable</td>
<td>* Subject to sample of learners used</td>
<td>* Estimation of item difficulty level when item calibration is not possible</td>
<td>Johns et al. 2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prior in Bayesian estimation algorithm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learner's feedback</td>
<td>* Always estimable</td>
<td>* Interrupts learning process</td>
<td>* Estimation of item difficulty level of new items</td>
<td>Chen et al. 2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* Time consuming</td>
<td>Prior in Bayesian estimation algorithm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expert's feedback</td>
<td>* Always estimable</td>
<td>* Time consuming</td>
<td>* Estimation of item difficulty level of new items</td>
<td>Linacre 2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prior in Bayesian estimation algorithm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paired comparison</td>
<td>* Always estimable</td>
<td>* Interrupts learning process</td>
<td>* Estimation of item difficulty level of new items</td>
<td>Ozaki &amp; Toyoda 2006, 2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* Time consuming</td>
<td>Prior in Bayesian estimation algorithm</td>
<td></td>
</tr>
<tr>
<td>Ability level</td>
<td>Weighted adaptation based on background characteristics</td>
<td></td>
<td>* Indication of learner's ability level</td>
<td>* Loss of some information</td>
<td>* Ability level estimation on the basis of few responses</td>
<td>Masthoff 2004</td>
</tr>
<tr>
<td>Alternative item selection algorithm</td>
<td></td>
<td></td>
<td>* More efficient than Fisher information when ability level is unknown</td>
<td>* More radical than weighted adaptation based on background characteristics</td>
<td>* Ability level estimation on the basis of few responses</td>
<td>Chen et al. 2000</td>
</tr>
<tr>
<td>Elo rating systems</td>
<td></td>
<td></td>
<td>* Perception of atypical learning patterns</td>
<td></td>
<td>Tracking of ability level in item-based learning environments</td>
<td>Brinkhuis &amp; Maris 2009</td>
</tr>
</tbody>
</table>
paired comparison are less applicable for an entire item bank calibration as it is time-consuming, and it can interrupt the learning process. Nevertheless, all the mentioned alternative estimation methods can be applied for item difficulty estimation of new items. Furthermore, we can implement the item difficulty values that are obtained with one of those alternative estimation methods into the IRT-based estimation algorithm, yielding a faster acquisition of reliable item difficulty estimates (Swaminathan et al. 2003). The algorithm to estimate the learner’s ability level faces two problems: the cold start problem and the problem of the evolving ability level. The former problem can be solved by weighted adaptation based on background characteristics and a modified item selection algorithm. The latter problem can be solved by implementing the Elo rating system to track the learner’s ability. The advantages of the Elo rating system are that it can quickly follow the change in knowledge level and that it can easily be extended. An example of this last advantage is the incorporation of background characteristics. When the learner starts working in the learning environment, a realistic starting value of the learner’s knowledge level based on background characteristics can enhance the estimation process in the Elo rating system. The problem of the item selection algorithm broaches the question of what item selection criterion should be implemented and of whether learners should have some control over it. The first question focuses on whether to select items that maximize the measurement precision and thereby leaving the learner with a probability of 50% to answer an item correctly or whether to select more easy or difficult items. The answer to this question should be explored into more detail in further experimental research. The second question concerns the learner’s control over the item selection algorithm. More specifically, should the learner be provided the freedom to choose the item difficulty level of the next item or not? Studies in CAT have shown that giving a person control leads to higher motivation. However, because the persons do not always select the item difficulty that matches their current ability estimate, the measurement precision is biased. We could combine this research result with the research results regarding the modest impact of using psychometrically non-optimal items on the measurement precision. This could yield an item-based ITS where learners have the possibility to select the item difficulty level of the next item out of a specific item difficulty range. This difficulty range does not lead to a high impact on the measurement precision. In such a situation, the learners perceive that they have some control over the learning environment, which might lead to higher motivation.

We can conclude that IRT is potentially a valuable method to adapt the item sequencing in item-based ITSs to the learner’s knowledge level. Applying IRT to learning environments, however, is not straightforward. In this paper, we wanted to identify associated problems and suggested solutions based on the literature on testing and learning environments, whether they were IRT-based or not. Further research in which these ideas are empirically evaluated is required.

References


