Tracking educational progress
Brinkhuis, M.J.S.

Citation for published version (APA):
Chapter 1

Measuring Change: Opportunities to Learn

1.1 Introduction

Traditionally, educational data is collected from an entire classroom of students taking the same test at the same time, using paper and pencil. However, particularly from the perspective assessment for learning (AFL) (Bennett, 2011; Black & Wiliam, 2003), there is interest in closely measuring progress and the individual development of ability (Wiliam, 2011). Knowledge of students’ current ability levels facilitates the possibility of tailoring education to individual needs, providing all students with opportunities to learn at their individual level. The availability of up-to-date ability estimates requires frequent assessment of all students, with results available to both teachers and students in real time. Though not developed in his primer on computer adaptive testing (CAT), Wainer (2000, p. xi) predicted that computerized measurement would serve such new purposes: “These purposes will include continuous measurement of dynamically changing quantities in educational settings, and other purposes not possible with static paper-and-pencil delivery and static metaphors of measurement”.

Examples of systems that involve continual measurements and dynamic changes, are electronic learning environments. A discussion of several types of such environments has been provided by Wauters, Desmet, and Van den Noortgate (2010). Examples include computer adaptive learning (CAL) system (e.g., Eggen, 2012; Veldkamp, Matteucci, & Eggen, 2011; Bloomfield, Roberts, & While, 2010; Ecalle, Magnan, & Calmus, 2009), intelligent tutoring systems (ITS) and adaptive hypermedia systems (e.g., Brusilovsky, 1999, 2001) and computer adaptive practice (CAP) systems (Klinkenberg, te meier, & van der Maas, 2011). Techniques related to CAT (e.g., Weiss, 1982; Wainer, 2000; van der Linden
& Glas, 2002; Eggen, 2004) are sometimes used in such systems to facilitate adaptive item sequencing. Though CAT on the one hand and CAL and CAP on the other are certainly related, their focus is different. Whereas the focus of CAT is to efficiently estimate abilities of test takers, CAL and CAP systems aim to provide learners with (practice) items tailored to their current ability level. See for example Ericsson, Krampe, and Tesch-Römer (1993) or Ericsson (2006) for distinctions between practice and learning.

Education data gathered over longer periods of time poses some interesting psychometric challenges. Quite generally it might be that measurement itself induces change, but more specifically, we expect that data gathered from systems where learning is one of the explicit goals, e.g., through practice and feedback, indeed involve dynamically changing model parameters and intricate non-linear dynamics. For example, differential development of abilities is expected due to instructional sensitivity and educational reform (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002), effects such as the pygmalion effect (Rosenthal & Jacobson, 1968) and methods of instruction (Bloom, 1984).

In the following section, we identify two problems with psychometric models in relation to CAP and CAL systems, which are more generally also applicable to ITS. First, we focus on the assumption of parameter invariance, which is quite commonly made in psychometric models. Especially in CAP and CAL systems, invariance of the involved pupil ability parameters and item difficulty parameters, is both unlikely and undesirable, e.g., given that practice should cause ability to increase. Second, we discuss that the explicit modeling of the development of ability may not be feasible, since not much is known on how development might take place under the influence of practice and feedback.

1.1.1 Common psychometric models

The notion that certain abilities or traits can be considered as fixed is deeply rooted in educational measurement, specifically in item response theory (IRT). Though Lord and Novick note in their seminal work on psychometrics “ability does not change” (1968, p. 13) and “[the true score of a person] is a constant” (1968, p. 30), usually, it is implicitly assumed in their work, and in educational measurement in general, that quantities such as one’s ability do not change, at least not during testing. In the same year, the promises for measuring progress using computerized item banks are made explicit: “The item bank could provide […] a single scale for measuring progress throughout secondary school—and in principle there is no reason why the scale should not be extended in both directions” (Choppin, 1968, p. 872). The idea of an invariant measurement
scale is expressed explicitly here. Apparently, such a measurement scale, and the characteristics of the items underlying such a scale, were considered to be unchangeable by feedback, learning, or intricate non-linear dynamics taking place in measurements over longer periods of time.

While the assumption that such effects do not take place might be reasonable and pragmatic for many situations, it is our contention that, in an increasing number of applications such as CAP and CAL, this assumption no longer holds. Consequently, many of the statistical methods that are commonplace in educational measurement are not suitable for these contexts. Though it is easily asserted that the ability of students changes through time, it is likely that other concepts also change, such as item difficulty and the measured construct. The statistical models that are currently used, such as classical test theory and IRT (Lord & Novick, 1968), and the inferential tools associated with them, e.g., maximum likelihood (ML) estimation, are not appropriate for this type of data owing to their assumptions that parameters are time invariant (e.g., Kane, 2011, p. 13). The common use of models that implicitly assume parameter invariance on data that might not support this assumption seems to be a typical case of the law of the instrument (Kaplan, 2004), and was already expressed as a concern by Roid (1986, p. 34): “...the indiscriminant application of IRT models to any tests, especially those not specifically designed to fit the model, would seem to be ‘technology gone wild’.”

The explicit modeling of change or growth does not seem to be a definite solution to the problem of changing parameters. In a context with continual changing model parameters such as item difficulties and person abilities, methods that are suitable for comparing ability change within a test (Verhelst & Glas, 1993; Verguts & De Boeck, 2000; Rijkes & Kelderman, 2007; Finkelman, Weiss, & Kim-Kang, 2010), or between a few tests (Fischer, 1989; Embretson, 1991; Keuning & Verhoeven, 2008), are not generally applicable.

Though in many educational contexts little is known about what growth patterns to expect, cf. McArthur and Choppin (1984), many models have been developed to deal with possibly complex developmental patterns, see for example the collection by Collins and Sayer (2001). However, the principle problem with modeling change addressed in this thesis, is that we expect intricate non-linear dynamics due to both feedback and the possibility that measuring the mind changes the mind, both important parts of CAL and CAP systems.
1.1.2 Alternatives

To identify statistical tools that deal with the type of data described in this introduction, we will first look at two different fields of research where frequent large-scale data processing takes place, routinely reporting results in real time on parameters that change over time. We will describe and evaluate the possible uses of techniques from these fields for educational measurement. We will then discuss several alternative approaches, designed to utilize the strengths but resolves the weaknesses of the existing methods, and hence be more suitable for this field of research.

1.2 Measuring change

1.2.1 Sports

Sports is a field in which we typically deal with changing abilities, with measurements over longer periods of time. As a specific example, we consider chess playing, were a tradition has developed to measure the changing abilities of chess players over time. Though apparently unrelated to educational measurement, certain developments in the rating of chess players are of interest. As early as the beginning of the previous century, there were developments in rating chess players which resulted in the formulation of one of the very first latent trait models by Zermelo (1929), who proposed a model that is now known as the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959), see Glickman (2013) for a historical perspective. In the early 1960s, Elo developed a method for tracking individual and idiosyncratic changes in chess playing proficiency over long time periods in which individual players played at irregular intervals with irregular frequency (Elo, 1978; Batchelder & Bershad, 1979). Currently, many chess federations, such as the The United States Chess Federation (USCF) and the World Chess Federation (FIDE), use the Elo rating system (ERS), or an adaption thereof, to rate their players.

The rating of chess players is of interest because chess can be regarded as a large-scale adaptive test, with frequent measurements (van der Maas & Wagenmakers, 2005). It is adaptive since chess players participate in competitions of players with similar playing strengths. When the ability of a player increases, this player will likely participate in other competitions that are more suitable to his or her playing strength. In addition, measurements occur quite frequent because chess players usually play with some regularity. The ERS uses quite simple calculations to provide an updated rating after every match, where the
rating difference between a pair of players can be used to calculate winning probabilities. The ERS has been used for over half a century in a sport where the stakes are considered high, providing it with some external justification.

The ERS, specifically Elo’s *Current Rating Formula for Continuous Measurement*, is introduced as follows (Elo, 1978, p. 25):

\[ R_n = R_o + K(W - W_e) \]  

(1.1)

- \( R_n \) is the new rating after the event.
- \( R_o \) is the pre-event rating.
- \( K \) is the rating point value of a single game score.
- \( W \) is the actual game score, each win counting 1, each draw 1/2.
- \( W_e \) is the expected game score based on \( R_o \).

In a match between two players, each of the players has a rating \( R_o \) before a match. The rating \( R_n \) is increased if he or she wins and decreased in the case of a loss. The amount of increase or decrease in ratings points depends on their relative difference in ratings before the match, expressed as the winning probability \( W_e \) of the player. If the player is expected to win, and he or she does, few points are gained. However, if someone wins against the odds, more points are gained by the winner. Though Elo (1978) modeled the expected winning probability \( W_e \) using the normal distribution, as in Thurstone’s Case V model (Thurstone, 1927), in the same work he discusses the possibilities of using the Verhulst or logistic function, cf. the BTL model (Bradley & Terry, 1952; Luce, 1959) or the Rasch model (RM) (Rasch, 1960). Clearly, other choices for \( W_e \) are possible, e.g., involving response times (Maris & van der Maas, 2012). The factor \( K \) determines the step size of the algorithm, where a large step size allow the ratings to adapt quickly to changing abilities, at the cost of additional noise.

Applications of the ERS are numerous, especially in the rating of sports and games (Fahrmeir & Tutz, 1994; Coulom, 2008). In addition to the use of the ERS to rate pair-wise comparisons of persons over time, one can also rate other pair-wise comparisons, such as a person answering an item, resulting in either a correct response (a win), or an incorrect response (a loss). This conceptual extension is made in two state-of-the-art applications of AFL that rate chess tactics, simultaneously rating both players and chess tactics items. Both applications use an extension of the ERS developed by Glickman (1999, 2001) to deal with rating uncertainty over time. Each player receives an immediate update of his rating after each response to an item, using a combination
of the correctness of a response and the response time. Both applications are websites for learning chess tactics. Both Chess Tempo (CT)\(^1\) and Chess Tactics Server (CTS)\(^2\) provide large-scale CAP environments, consisting of a large item bank and many players. The items generally consists of specific board positions followed by a move, after which the player has to decide what the best following move is. The CTS item bank consists of about 37 thousand items, and over 85 million chess problems have been served to about 65 thousand players. The CT item bank consists of over 55 thousand items, and over 207 million chess problems have been served to about 310 thousand players (numbers from April, 2013).

Applications of the ERS are also found in educational measurement, such as in training and learning environments (Wauters et al., 2010). Klinkenberg et al. (2011) discuss a novel CAP and progress-monitoring system for arithmetic skills called the Math Garden\(^3\). In the Math Garden, individual ability level is estimated in real-time, while an individual answers math problems. Item are selected using a predetermined success rate, which is based on the estimated ability of individuals (Jansen et al., 2013). The Math Garden uses an extension to the ERS to include response times, as described by Maris and van der Maas (2012). In Math Garden, more than 500 thousand math problems are solved on a daily basis, involving over 30 thousand students from 900 schools and cumulating to over 220 million records in total (numbers from April, 2013). Item banks of this magnitude are quite unique in educational measurement applications.

More exotic applications of the ERS include the rating of looks (Schwartz, 2007), animal dominance (Albers & de Vries, 2001), programmers (Forišek, 2009) and plant breeding (Simko & Pechenick, 2010). Many extensions and competing rating systems are developed for specific goals, e.g., Microsoft’s TrueSkill algorithm to rate both players as well as groups of players in large scale on-line games (Herbrich, Minka, & Graepel, 2006; Dangauthier, Herbrich, Minka, & Graepel, 2007).

The large size of the databases of CT, CTS and Math Garden, and the use of real-time rating updates, disqualify estimation methods commonly used in educational measurement, such as ML estimation, due to computational limits. On the other hand, rating systems such as Elo’s are tractable, computationally simple and easily implemented. It is a valuable tool which provides ratings of both players and items in large-scale applications of CAT and CAP. However,

\(^1\)http://chesstempo.com
\(^2\)http://chess.emrald.net
\(^3\)http://www.mathsgarden.com
there is a challenge in using the ERS in educational measurement. In a traditional testing framework, we ideally let students take sufficiently long and reliable tests so that their scores are an accurate estimate of their ability. Since the reliability of a single-item response is low, the frequent rating updates in rating systems such as the ERS cause the ratings to be quite noisy and unstable. The key problem with the ERS is the fact that the distribution of noise is not well defined. Without a known distribution for ratings, no straightforward methods are available to evaluate the reliability of ability estimates from rating systems like Elo’s.

1.2.2 Engineering

Another field in which we see data being gathered at a high frequency, and involving changing parameters, is engineering. We focus on a specific application, the Global Positioning System (GPS). GPS is a widely used technique to determine location. Navigation devices in cars, planes, or in mobile phones use signals from a number of satellites to provide an estimated position on the world map. This problem is related to educational measurement in that we want to track a student’s position in a certain ability space, just as our navigation device tracks our location on the globe; a metaphor also found in Wainer (2000, p. xi). The GPS data obtained is similar to data from CAP systems in that a GPS capable device tracks one’s continually changing position using somewhat noisy measurements, gathered on a regular basis. We can also regard responses to items as noisy estimates of someone’s changing ability, where we try to pinpoint someone’s ability on our ability map. This makes the statistical techniques used in GPS promising to look into.

One of the statistical techniques required to perform the calculations in GPS is the Kalman filter (KF) (Kalman, 1960; Welch & Bishop, 1995; Arulampalam et al., 2002). KFs are a popular statistical technique with applications in, amongst others, economics, dynamic positioning, chemistry and speech recognition. The specific application of these filters to global positioning and the use of extra information like inertia sensors is described in many engineering textbooks, e.g., Grewal, Weill, and Andrews (2001). A KF typically consists of two models. The first being a measurement model, relating noisy measurements to an estimate of the current state (e.g., position on the map or ability level). The second being a system model, describing the change of the state over time. The system model in GPS can be based on other sensor input to provide information on where someone is going, for example, data from an accelerometer. If someone accelerates in a certain direction, the noisy satellite
position estimate can be enhanced by this information. The basic KF is a two-step procedure with a prediction and an update step. In the prediction step, the system model is used to predict the next state. In GPS, this would be the acceleration in a certain direction predicting your next position. In the update step, Bayes theorem is used to combine knowledge on one’s predicted position with the possibly noisy GPS position from the measurement model. Though not introduced as such, KFs can be described as specific and tractable forms of sequential Bayesian algorithms. The KF makes several specific assumptions, such as Gaussian distributed states and noise, and linear regressions in the systems model and in the measurement model. If these assumptions hold, the KF is a tractable and optimal algorithm for recursively calculating the posterior density.

The application of a recursive Bayesian algorithm, such as the KF, to educational measurement has several advantages over the ERS. It has several statistical properties concerning the distribution of estimates, which can be used for testing for change, checking model fit, and more. However, there are also downsides. It requires the same two models as in the GPS application. The measurement model can be some familiar item response model which is readily available. The system model facilitates prediction, which is a growth model in educational measurement. While rather complicated growth models can be specified (e.g., Speekenbrink & Shanks, 2010), theories on how growth develops throughout education are not readily available. This complicates a straightforward implementation of recursive Bayesian algorithms in student monitoring systems.

In addition, the KF has some quite strong assumptions, such as a Gaussian distributed posterior density at every time step, and both the system and the update model are linear functions of their parameters and include Gaussian distributed independent and identically distributed (i.i.d.) errors. To relax these assumptions one can use extensions such as particle filters. Though there are many examples of the use of such models in educational measurement (Fahrmeir & Tutz, 1994; Molenaar, 1987, 2004; Hamaker et al., 2005; van Rijn, 2008; Wang et al., 2013), the rather strict assumptions of the KF or the quite intensive calculations of particle filters make these filters less suited for providing results in real time in large scale educational measurement contexts. Moreover, it has already been mentioned that measuring someone’s mind is likely to change it, especially through practice and feedback, which would require changes to the measurement and system model too. The result is a intricate non-linear dynamic system, for which filters with fixed models might not be suitable.
1.2.3 New tools for measuring change

There are several strengths to the two techniques that have been discussed in relation to educational measurement. The application of the ERS in chess ability estimation and, in particular, in the CTS, CT, and the Math Garden, demonstrates that tracking ability in a testing context with many persons, items, and frequent responses is feasible. A weakness of the system is a lack of control of the distribution of ability estimates, which is crucial for educational measurement applications. For example, without defined distributions of abilities, it is challenging to interpret ability development since there is no clear indicator of what is growth and what is noise. The application of GPS uses KFs, i.e., sequential Bayesian filters. The strength of this approach is that we can accurately track changing abilities with full control of the distributions involved. A weakness of applying such filters for tracking ability, is that growth model are assumed to predict ability change, while information on how such a growth model should look is not readily available and moreover, subject to change due to feedback.

In constructing algorithms for measuring change in this thesis, we require some specific features for our educational measurement applications. First, such algorithms must be able to closely follow the abilities of individuals over time, and therefore to update estimates after each response, as the ERS allows for. Second, we require that our ability estimates be distributed in a specific manner around the student’s true ability level, as in applications of the KF. Third, we want to refrain from requiring a specific growth model for the development abilities, as these are often not readily available in applications in educational measurement.

1.3 Thesis chapters

In chapter 2: Tracking, a definition and formalization of psychometric trackers is presented. Trackers are defined as dynamic parameter estimates with specific properties. As indicated the previous section, these properties include the ability to adapt to parameter changes, and a specified noise distribution. Chapter 3: Detecting differential development, presents a specific tracker suitable for detecting differential development over time between pairs of items or pairs of persons. The tracker is designed to work on data that is filtered from large data collections, extracting only relevant information on differences between pairs of items or pairs of persons. Since this tracker generates i.i.d. Bernoulli trials under the RM, one can several statistical tools to test for differential development.
Chapter 4: Metropolis ability trackers, introduces trackers using the Metropolis algorithm. Two trackers are worked out, both using item responses directly in the Metropolis algorithm, i.e., using item responses as either proposal values or to determine acceptance or rejection. In the last chapter, chapter 5: Dynamic estimation in the extended Rasch model, an extended marginal Rasch model (ERM) with a Bayesian estimation procedure is presented. It allows for parameter estimation in static contexts using fully observed data matrices and in sequential adaptive contexts, such as CAT, where model parameters are estimated as data becomes sequentially available. In dynamic contexts, such as CAL or CAP environments where model parameters can change over time, it can be used to simultaneously track the development of item and person parameters.

References


Chapter 1. Measuring change


