# Vertical and Stationary Scales for Progress Maps

Russell G. Almond
Florida State University
Educational Psychology and
Learning Systems
Tallahassee, FL
ralmond@fsu.edu

Ilya Goldin
Center for Digital Data,
Analytics, and Adaptive
Learning
Pearson Education
ilya.goldin@pearson.com

Yuhua Guo & Nan Wang
Florida State University
Educational Psychology and
Learning Systems
Tallahassee, FL
[yg07c,nw13]@my.fsu.edu

## ABSTRACT

Students and instructors would benefit from a graphical display of student proficiency throughout a course. However, valid and reliable proficiency estimates based on modern statistical techniques require data that are not usually collected in traditional instruction. For example, problems that students solve on tests and homeworks may not be properly equated to a vertical scale; building a true vertical scale requires that overlapping anchor items be administered in way that supports the estimation of student growth between assignments. This paper suggests an alternative, a *stationary* scale in which the expected student growth is subtracted out so that a student making normal progress remains at the zero point in the scale. We define the stationary scale model and validate it on a real-world data set of student answers to homework items. We further produce a Progress Map, a visualization of student proficiency throughout a course.

## Keywords

Ability estimation, Homework, Item Response Theory, Kalman Filter, Smoothing, Partially Observed Markov Decision Processes, Progress Maps, Graphical Displays, Vertical Scales

## 1. INTRODUCTION

All students in a course usually have the question,[1] "Am I on track to master the objectives listed in the course syllabus?" Good students will revisit this question throughout the course and adjust their studying strategy if they are at risk of not mastering the objectives. Instructors have two related questions: "Are my students (as a class) on track to master the objectives?" Again, good instructors will monitor the answers to these questions and "Which students are at risk to not master the objectives?" and adjust the instruction as needed. This is an issue of measurement.

Optimal measurement is not the only consideration when instructors choose items for assignments and quizzes. Instructors primarily choose items to practice objectives recently introduced in class. Although using multiple items per instructional objective produces more reliable measurement, the instructor must balance the test length with student fatigue. If students do not all work on the same items, as may be the case if items are adaptively selected or automatically generated by an Interactive Learning Environment (ILE), then despite item differences, instructors rarely attempt to put the scores onto a common scale. In particular, ensuring that there are sufficient overlapping items between forms to do any kind of common item equating is usually a low priority when choosing items for an assignment.

The lack of a proper equating design in the assignments complicates estimating student *growth* over time from homework assignments. Most procedures for estimating growth require all of the assignments to be linked to a common *vertical scale* [13]. One method of constructing a vertical scale requires overlapping items between adjacent assignments. This is a problem for the instructor because the overlapping items will either cover problems from prior or future topics. As the instructor's goal is maximizing the time spent practicing the current topics, such review and preview items are seldom included on assignments. A second method of constructing a vertical scale requires administering items from throughout the course at a common time point to a group of students who have been subject to a standard set of instruction. Usually courses offer limited opportunities (e.g., pretest, midterm, final) to do such calibration.

To address the problem of tracking student progress across time using homework assignments that have not been vertically scaled, this paper introduces an alternative to the vertical scale called a *stationary scale*. The basic assumption of the stationary scale is that the average student ability and the assignment difficulty increases at the same rate; in other words, the instructor designs each assignment to match the expected ability distribution of the students at the time when the assignment is due. On the stationary scale, the expected ability of a student who is on track remains at 0 throughout the course. This is equivalent to a stationary time series. We use the stationary scale to construct a unified model for multiple assignments over time, incorporating at once all the observations about each student over the duration of a course. This has two benefits: First, it lowers the error of measurement for each assignment

[1] The actual question is something more like "Will I get a good grade?" Good grades, however, should follow from mastering the objectives.

and each student. Second, it puts the assignments on a common scale so that growth can be interpreted as deviations from expected growth.

The next section lays out a time series framework for assignments and talks about previously developed models for growth and observed outcomes on assignments. The following section describes the stationary model and the calibration of models to the stationary scale. The fourth section explores the application of the stationary model to a database of online homework results. The last two sections explore some possible graphical displays and offer suggestions for improvement and future work.

## 2. MODELS FOR STUDENT CHANGE OVER TIME

Figure 1 shows a general Markov decision process framework for integrating information from multiple assignments over time [2]. Here $S_t$ represents the latent ability of the student, and $O_t$ represents the observed outcomes of the assignment offered at Time $t$ (both of these quantities could be multidimensional). The nodes marked *Activity* represents what activity the instructor chooses for the students between sessions. If the problem is to choose an optimal strategy for selecting activities, then Figure 1 is a partially observed Markov decision process (POMDP)[3]. To simplify the problem, assume that all students get the same action, "continue with the next lesson according to the syllabus," at each time point. This reduces the problem from a POMDP to a hidden Markov model (HMM).

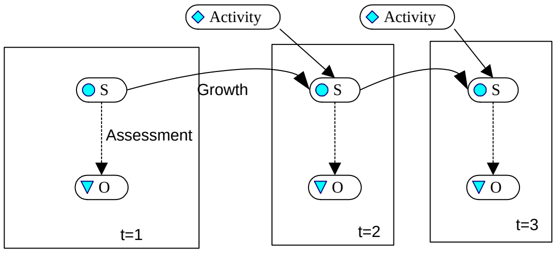

**Figure 1: Generic Framework for Accumulating Assessments over Time [2].**

The POMDP/HMM framework decomposes the modeling problem into two pieces: what happens within a single vertical time slice (i.e., within an assignment) and what happens between time slices (i.e., the growth model). In a single time slice, familiar models such as item response theory (IRT) apply. For the growth model, the Brownian motion process provides a simple starting point. The biggest drawback of this framework is that the scale of the latent variable, $S_t$ is not identified; this causes difficulty in estimating model parameters from data [1]. The common solution is to put the latent variables onto a vertical scale.

### 2.1 IRT Models for observed outcomes

For simplicity, assume that the latent state of Student $i$ at Time $t$, $S_{it}$, can be represented with a unidimensional, continuous random variable. If every student is in the same course with the same syllabus, the distance along the common path through the multidimensional space defined by the course syllabus that the student has traveled [14] can appear like a single dimension. Care must be taken when comparing the stationary estimates from different courses as different paths through the multidimensional space will give different meaning to the same value on the unidimensional scale.

Again for simplicity, assume that observable outcome from Student $i$ interacting with the assignment given at Time $t$ is a binary vector $\mathbf{O}_{it} = \{O_{i1t}, ..., O_{iJt}$, where $O_{ijt} = 1$ if Student $i$ got a correct answer to the $j$th item given at Time $t$, and zero otherwise. Item response theory (IRT) models the likelihood of each observation as conditionally independent given the latent state of the student [11]. There are several possible models; this paper uses a one-parameter logistic (1PL) or Rasch model:

$$P(O_{ijt} = 1|S_{it}) = \text{logit}^{-1}(S_{it} - b_{jt}) \quad (1)$$

$b_{jt}$ is an item-specific difficulty parameter.

In an IRT model, the scale and location of the latent variable $S_{it}$ is not identifiable from the data. One conventional way to resolve this, as we do here, is to set the average of the difficulties, $b_{jt}$, to 0. If we calibrate the IRT model using the data from a specific class taking a specific assignment, then this will produce a set of *class-specific IRT parameters*.

IRT has mostly been applied in the world of high-stakes testing, where each examinee attempts a problem exactly once. However, in online learning, most of the assignments will be homework and other lower-stakes assessments. Online homework systems frequently allow multiple attempts at an item, and access learning aids (e.g., tutorials and worked solutions to similar problems) when difficulty arises. Furthermore, the course policy may allow multiple attempts at the whole assignment, even for tests and quizzes which only allow a single attempt at each item within the assignment.

In the world of online homework systems that allow multiple attempts it becomes difficult to define "correct". Two possible definitions are *Correct-on-first-try*—solving the item on the first item-level attempt in the first assignment-level attempt without the use of learning aids,—and *Eventually-correct*—solving the item on any attempt with or without learning aids. The two different definitions of correct will give slightly different meanings of proficiency. These correspond to Falmagne's *inner* and *outer fringes* of the learning space [8]): correct-on-first-try corresponds to the inner fringe—those things that the student can do without assistance,—and eventually-correct corresponds to the outer fringe—those things the student can do with assistance.

### 2.2 Brownian Motion Growth Model

As the goal of this paper is to produce quick estimates of proficiencies that can be used to track student progress, the simplest growth model discussed in [2] provides a good starting point. Assume that between Times $t-1$ and $t$, the average expected growth following the syllabus is $\delta_t$, and let $\Delta t$ be the time (either in calendar time, or some measure of progress through the course as number of chapters of the textbook covered) between the assignments at Times $t-1$ and $t$. Further assume that the growth for an individual student over one time period is normally distributed around the class average, let $\omega^2$ be the variance over a unit time interval. The variance of the growth between two assignments is $\omega^2 \Delta t$, that is, the variance of the growth is proportional to the elapsed time. Thus, $S_{it} \sim N(S_{i,t-1} + \delta_t, \omega^2 \Delta t)$. This is a non-stationary Brownian motion process.

The Brownian motion model implies that the less frequently the student is assessed, the more uncertainty there is about the student's ability. Almond [2] suggests that when the variance of the growth increments, $\omega^2 \Delta t$, is small with respect to the standard error of measurement, $\tau_t$, the ability estimates can be smoothed across time to have a lower mean squared error. Almond suggests a number of techniques for smoothing: a simple exponential filter (down weighting each prior observation by a factor $\lambda$), the Kalman filter [10] (this assumes that $O_{it}$ is approximately normally distributed given the latent ability) and the particle filter [7] (which supports many models for both within and between time slices). These models can also forecast future ability states.

Attempting to estimate the within-time slice (observable outcome) and between time-slices (growth) models at the same time can cause difficulty [1]. In particular, either the average ability increase $\delta_t$ or the average difficulty of the items at Time $t$ cannot be identified from the data. The usual approach in these circumstances is to put the assessments given on each time slice onto a vertical scale.

## 2.3 Vertical Scales

Educators frequently want to measure student learning using tests administered at different times and covering different but overlapping content. In order for differences in the test scores to be meaningful, the tests must be placed onto a common or *vertical* scale [13]. Vertical scales can be challenging to develop [15] and require difficult to verify assumptions [9].

In particular, constructing a vertical scale usually requires *anchor items*, items that are placed in two adjacent tests in the series. By assuming the anchor items have the same psychometric properties in both administrations (an assumption which is open to question [12]) the adjacent test forms can be equated. While anchor items are included in high-stakes testing programs, they are seldom included in homework assignments, where the focus is maximizing practice of the most recently introduced material.

An alternative is to place the anchor items into a separate test which is administered at a single time, so the students see the items covering many parts of the curriculum at the same time. Pretests and final exams provide natural experiments of this type. Even so, the number and quality of the anchor items controls the quality of vertical scale (in particular, the standard error of the equating that underlies its development). It is unusual to have enough anchor items to properly build a vertical scale for homework data.

Assume that a number of anchor items have been assigned difficulty parameters on the vertical scale. Let $J_t$ be the subset of items in the assignment given at Time $t$ that have associated difficulty parameters on the vertical scale, $b^*_j$. To equate the current assignment to the scale defined by the vertical scale [11], replace the difficulties in the assignment, $b_{jt}$ with the equated difficulties:

$$b''_{jt} = b_{jt} - c_t \ , \tag{2}$$

where

$$c_t = \sum_{j \in J_t} b_{jt} - b^*_j \ .$$

Note that the quality (i.e., standard error) of this equating will depend on the number of anchor times in each assignment. As homework assignments are typically short, the number of available anchor items is typically small and hence the quality of the vertical scale is questionable.

# 3. STATIONARY SCALES

Consider the problem of estimating students' abilities as they progress through a course. Assume that homework, quizzes and tests are given online, so that the course learning management system has a record of each item from each assignment, as well as which items were and were not attempted. Furthermore, assume that the class size is large enough that parameter estimates from calibrating the IRT model given in Equation 1 will have reasonable standard errors. The instructor would like proficiency estimates for each student at the time of each assignment, as well as end-of-course forecasts.

If the item parameters are not already available, the difficulty of each item must be estimated from the course data. However, the instructors usually assign items when they make sense according to the syllabus of the course, i.e., shortly after the objectives covered in the item were covered in class. This pattern of item assignment does not usually produce the kind of overlap needed to support the construction of a vertical scale.

The alternative we are proposing is a *stationary scale*. Let $\overline{b_t}$ be the average difficulty of the items given at Time $t$ on the vertical scale. Then the stationary scale is defined by assuming $\delta_t = \overline{b_t} - \overline{b_{t-1}}$. In other words, the difficulty of the assignments grows at the same rate as the ability of the students. To put the item parameters and ability estimates on the stationary scale, set:

$$S^0_{it} = S_{it} - \sum_{s=1}^{t} \delta_t \ , \tag{3}$$

$$b^0_{ijt} = b_{ijt} - \sum_{s=1}^{t} \delta_t \ . \tag{4}$$

On this scale, the ability of a person moving through the course at the pace determined by the syllabus will be a stationary (zero mean) time series. In particular, the growth model of the previous section will be a *stationary* Brownian motion process, $S_{it} \sim N(S_{i,t-1}, \omega^2 \Delta t)$. Stationary time series are simpler to work with than non-stationary time series. In fact, many books on time series (e.g., [4]) recommend differencing the time series to make it stationary before analyzing the data. Similarly, many filtering techniques which could be used to smooth the observed ability estimates assume stationary time series.

The key assumption of the stationary scale is that the instructor, in the process of constructing the assignments, has taken care of the vertical scaling problem. In particular, we assume that the instructor is picking items so that the expected percent correct, hence the average difficulty, is approximately the same on each assignment. (The first author has found that after 2 or 3 times teaching an elementary statistics course, the median score on the midterm exam is usually close to the target score of 85%.)

Unfortunately, this assumption is difficult to test in practice. A convincing test would require a data collection design similar to the one required for a true vertical scale. The following section explores some more superficial checks which can be done using an arbitrary set of homework data from a large class. The remainder of this section looks at two operations which are now possible under the stationarity assumption: comparing a class to a database of similar classes, and smoothing ability estimates over time.

## 3.1 Classroom Level Estimates: Equating a class to a database

Assume that the IRT parameters have be calibrated using data from a single class, and that the scale has been identified by setting the average difficulty of each assignment to zero. Care must be taken in the interpretation, because a sudden drop in the average class ability could mean that the assignment was poorly designed rather than the class is not meeting expectations.

For many instructors, it would be useful to compare the performance of their class to similar classes: perhaps the same class offered in different years, or similar classes offered by other instructors. In particular, if we have a database of homework results from different classes using the same text, and if we assume that all of the instructors who use this database are also assigning items according to the stationary scale, then we can equate each class to the scale defined by the database using Equation 2.

Again, a version of the stationarity assumption allows a meaningful interpretation of item difficulties averaged across different courses. If we assume that each instructor introduces the item when it is instructionally relevant, that is when its difficulty matches the average student ability in the class, then the average difficulties across all classes in the database will also be on a stationary scale of sorts. This is a stronger version of the stationarity assumption than the single class version. If historical homework results are drawn from one textbook and pool of items across any instructors and syllabi, this will result in differences in which book sections are covered, their relative emphasis, and timing of delivery. Applying the stationarity assumption to the whole database assumes that the variability in the item difficulties produced by the variations in context are ignorable.

## 3.2 Smoothing the Ability Estimates

The real advantage of the stationary scale is that we can use the model of Figure 1 to smooth the proficiency estimates. Thus, at any time point the best estimate for a student ability is a weighted average of the student's ability estimate at the previous time point and the estimate from the current assignment. The weights depend on the relative size of the standard error of measurement for the assignment and the variance between time steps in the growth process [2]. Assuming that the growth model is normal process and that the ability estimates are normally distributed around the true ability allows the Kalman filter to be used to smooth the proficiency estimates. Although the observations are binary, the shape of the likelihood for the ability in the IRT model is approximately normal with a mean corresponding to the point estimate and a standard deviation corresponding to the standard error of measurement.

Implementing the Kalman filter requires knowledge of the variance of the growth increments, $\omega^2$. If the Brownian model model holds, then the variance of the ability variable should increase linearly with time. We estimated a Rasch model using a regression with a random student effect [6] for each assignment. This produces a variance for the student ability variable at each time point. The slope of the line for the ability variances regressed against time provides an estimate of $\omega^2$. This provides all of the necessary information to smooth the ability estimates using the Kalman filter.

Smoothing across time points should reduce the standard error of the ability estimates. In particular, the ability estimate at each time point will have a standard error that depends on both the direct evidence from the current assignment and the indirect evidence from the past history and the typical trajectory of student abilities. This will given a pattern of constant ability (on the stationary scale) and shrinking standard errors for students who are progressing normally. It helps answer one question instructors often have: when is a low assignment score a one-off fluke and when is it an indication of a problem which requires attention. In the former case, the filter will smooth the estimate towards the student's usual performance; in the latter case, the instructor will see the student trend line drifting away from the average trend of the class.

The biggest problems for instructors are not the students who progress normally, but the students who do not, especially students who do not complete assignments. Following the Brownian motion growth model, student ability will have an expected increase of $\delta_t$ on the vertical scale for each assignment, or on the stationary scale, expected ability will stay the same. The standard error of that estimate should increase. In particular, the variance of the estimate will be $\omega^2(t-t_0)+\sigma^2_{t_0}$, where $\sigma^2_{t_0}$ is the standard error of the ability estimate at the last time $t_0$ for which work is available for the student. If the student later returns to a normal completion pattern, the standard error will once again decrease and the proficiency estimate will track the student's ability. If the student continues to not submit assignments, the uncertainty will grow steadily larger.

# 4. MODEL VALIDATION

In practice, the stationary assumption is difficult to test: a rigorous test requires overlapping items in same kind of pattern as is used to construct a vertical scale. There are, however, three consequences of the model that we can test. First, if we separate the data into two pieces the mean ability on the stationary scale should change at the same rate for both groups. Second, the smoothed estimates of ability should have lower standard errors than estimates using data only from the most recent assignment. Third, the filter should produce reasonable predictions for the current assignment based on past assignments.

We fit the model to a data set of 578 students, spread across 9 sections, enrolled in an Intermediate Algebra course in the same semester. Students completed homework assignments online using the MathXL system[2]. There were 18 assignments taken from 3 chapters, Chapters, 3, 6 and 9 (with time elapsed between chapters). The assignments ranged in length from 12 to 62 scorable item parts[3], with most assignments having around 20. The number of students active in the course declined over time ranging from 567 attempting the first assignment to only 408 student attempting the penultimate assignment.

We randomly chose 10% of the students as test data and used the remaining students for training data. Using only the training data, we fit a Rasch model to all of the items in each assignment using a logistic regression with a random effect for student (ability estimate) and a fixed effect for items (difficulty) [6]. We then put the item parameters for that assignment onto the stationary scale by subtracting the average difficulty for each assignment. Once we had the final item parameters, we constructed expected a posteriori (EAP) estimates and the corresponding standard errors for the ability of each student in both the training and the test samples. Because the EAP estimates would later be combined with prior information about the student ability in a filter, a flat prior was used for the EAP estimates.

Figure 2 shows the average EAPs for the training and test samples. Note that the two estimates track each other closely. The correlation between them is $r = 0.92 (n = 18)$. While this does not prove that the stationary assumption holds, it does demonstrate that if it holds for the training sample, it holds for the test sample as well.

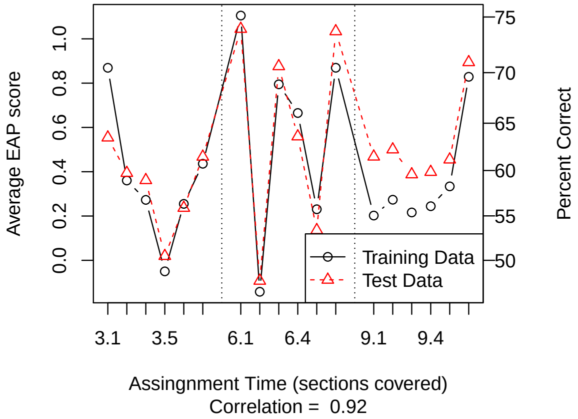

**Figure 2: Average EAP ability estimates over time**

If the Brownian motion model holds, the variance of ability estimates (from the random effects logistic regression) against time (sections of the book completed) should rise with a slope of $\omega^2$. Figure 3 shows the observed variances. For the first two chapters, where the variance is decreasing, the participation rate dropped by about 100 students. During the third chapter (Chapter 9) the sample size was more stable, and the population variance shows the expected linear increase. Consequently, we used the slope of the increase during the final chapter as the estimate of $\omega^2$.

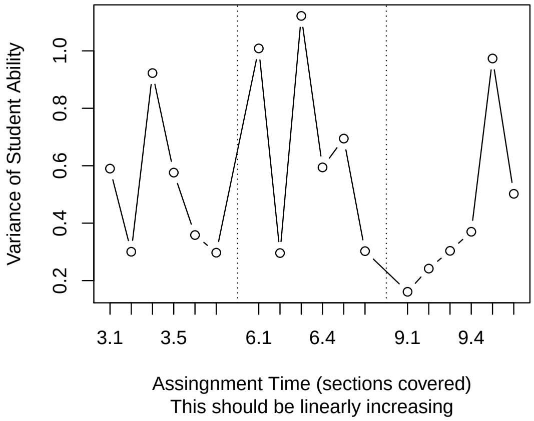

**Figure 3: Population ability variance over time**

Next, the ability estimates from the IRT calibration were smoothed using the Kalman filter. Because the smoothed estimates take both current and historical evidence into account, the standard errors for the smoothed estimates should be lower than that the standard errors of the original IRT estimates (which only include the current time point). Figure 4 verifies this is the case, plotting the average[4] standard error. The standard errors for the filtered estimates get better over time as the filter incorporates more data. Also, the standard errors increase during the intervals between chapters when some time elapses without measurements of student progress. It is also possible to run the filter backwards to get improved estimates for earlier time points (incorporating later data), but that was not done.

Validating the quality of the forecasts is difficult because there is no baseline to compare it against. As a weak form of validation, we look at the size of the average difference between the forecast from the filter and the EAP estimate from the IRT data (with no smoothing). Let $\hat{\theta}_{n,r}$ be the EAP estimate using only data from the current assignment at the time of Assignment $r$, and let $se(\hat{\theta}_{n,r})$ be its standard error. Further, let $\theta^*_{n,r}$ be the one step ahead forecast from the filter incorporating only data from past assignments. Let $z_{n,r} = (\theta^*_{n,r} - \hat{\theta}_{n,r})/se(\hat{\theta}_{n,r})$ be the standardized difference between the forecast and the current data only estimate.

Figure 5 shows the root mean squared standardized difference between the filter forecast and the IRT estimate using

[2] http://mathxl.com

[3] Some items had multiple scorable parts. We treated those parts as separate items, ignoring the dependence between the parts of the items.

[4] The average used here is the root mean square standard error, that is the square root of the mean prediction variance. Data points with missing assignment data are excluded.

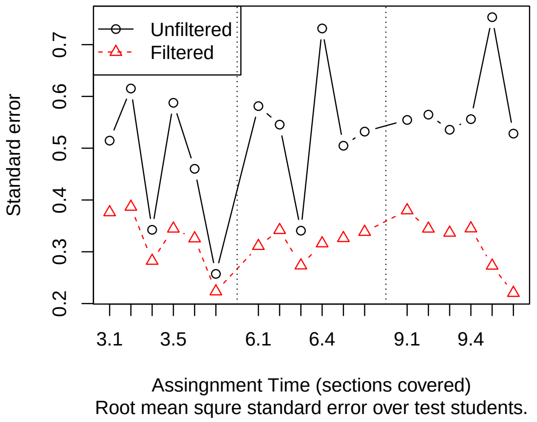

**Figure 4: Average standard error for test set**

only the current assignment data. The average (over the test set) difference is less than one standard error (of the IRT estimate) for all the time points, indicating the filter is doing reasonably well. Note, however, that the filter is doing fairly well even at time point zero where it is simply predicting the class mean for every student. Therefore, the positive result speaks more to the low information from the relatively short homework assignments than it does to the quality of the forecasts from the filter.

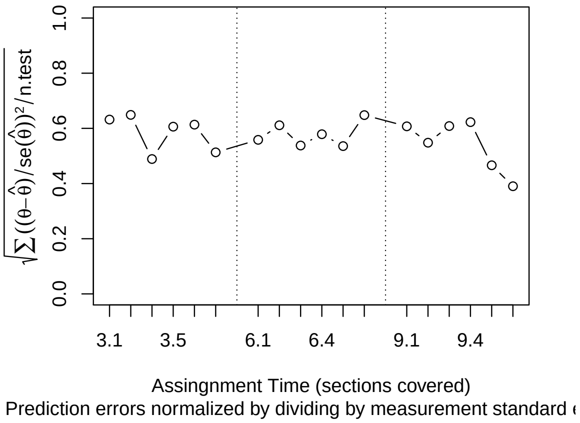

**Figure 5: Root Mean Squared prediction difference**

# 5. PROGRESS MAPS

To illustrate how the stationary scale can be displayed to an instructor or student, we display the progress of an arbitrarily chosen student. One simple graph plot the ability estimate from the IRT model against the time. We connect the measurements with a line; a dotted line indicates that one or more intermediate assignments is missing. There are a number of possible time scales to use. The method we found provided the clearest displays was to simply use the sequence number for each section, adding an extra step between the chapters. Figure 6 shows the result.

There are two additions we would like to make to this progress map. First, we would like the scale of the graph to give the visual impression that students making normal progress are increasing in proficiency. Second, we would like visual indications of the standard errors and expected performance standards.

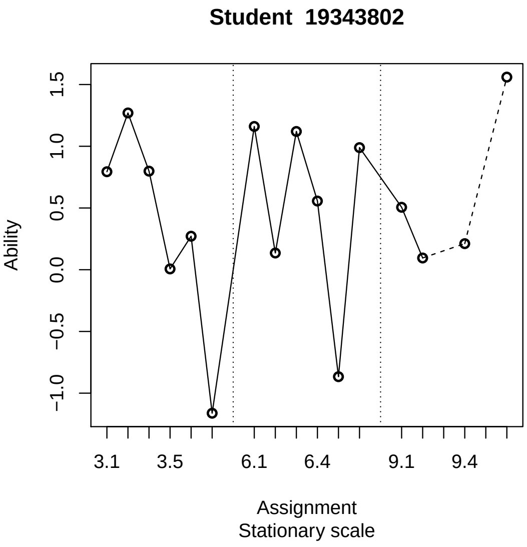

**Figure 6: Stationary Map: Graph of student progress on a stationary scale**

## 5.1 The Progress Scale

Although the stationary scale is mathematically convenient, the ability estimates for students making normal progress will remain flat. Students and instructors would prefer to see progress towards a goal, i.e., rising ability. If the ability increases, $\delta_t$, were known at each time point, then the data points could be put back on the vertical scale by inverting Equation 3 or 4. However, learning the ability increases is equivalent to the problem of learning the vertical scale. In particular, it requires the existence of a set of anchor items assigned in a pattern that supports the establishment of a vertical scale.

An alternative is to simply pick a convenient value, $d_t$, and set $\delta_t = d_t$. We call this scale the *progress scale*. There are now three possible scales (related through Equations 3 and 4):

**Vertical Scale** The values of $\delta_t$ are estimated from data. The quality of the scale will depend on the available anchor items.

**Stationary Scale** This defines $\delta_t \equiv 0$. Item parameters can be put on this scale by calibrating each assignment separately.

**Progress Scale** This defines $\delta_t \equiv d_t$ as an arbitrary series of constants. It can be readily produced from the stationary scale to make an increasing scale for display.

In our preliminary experience with graphical displays we have found letting $d_t = 1/K_t$, where $K_t$ is the number of sections in the chapter administered at time $t$ works well. This corresponds to an increase in one standard deviation in the population ability for each chapter covered in the text. Figure 7 shows an example. Here $d_t$ is set so that progress through 1 chapter is the equivalent of 1 point on the scale; the $d_t$ for a section is the corresponding fraction of the chapter. Note that the progress scale is slightly different scale from the section count scale used for the $x$-axis; that is why there appear to be sharp rises between the end of each chapter and the beginning of the next.

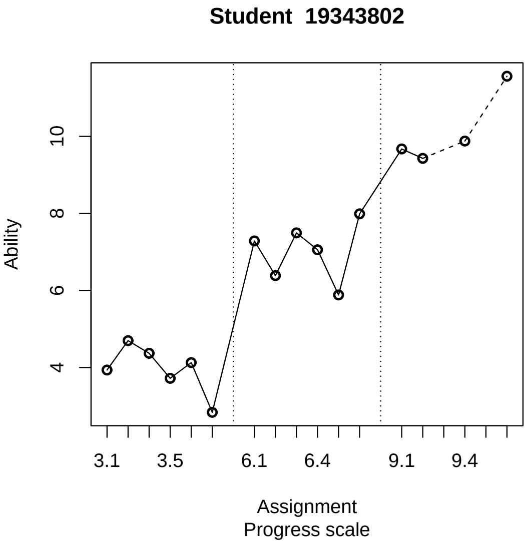

**Figure 7: Progress Map: Graph of student progress on a progress scale**

## 5.2 Error Bars and Control Limits

One drawback of the progress maps in Figure 6 and 7 is that they provide no information about the precision of the ability estimates. Figure 8 adds error bars to each point estimate extending ±2 standard errors from the point estimates. Both the point estimates and the standard errors are the unfiltered estimates from the IRT analysis.

Figure 8 also adds control limits to the display. The progressively darker shaded regions on the graph indicate areas of increasing concern for the student and instructor. When the point estimates pass the control limits, the size of the plotting symbol is changed to make the problematic data points more visible. The control limits are wavy instead of smooth because two different time scales are used, one (number of chapters completed) for adjusting the ability and one (assignment count) for plotting the time.

There are several ways of coming up with the control limits. Figure 8 uses a simple idea based on the IRT model. The instructor chooses a proportion correct for the assignment. To get the control limits, solve the IRT equation (Equation 1) for the ability that leads to that proportion for a zero difficulty item. This provides a roughly interpretable limit. The probabilities used in Figure 8 are .4, .2, .1, and .05.

The big advantage of the stationarity assumption is that it allows smoothing ability estimates using the Kalman filter. Figure 9 shows the filtered time series for the first student. Note that the error bars in Figure 9 are smaller than the error bars in Figure 8. This is an effect of the smoothing.

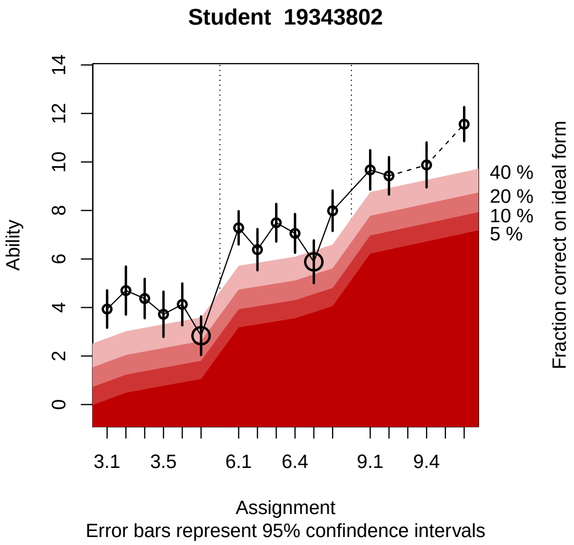

**Figure 8: Progress Map with error bars and fixed percentile limits**

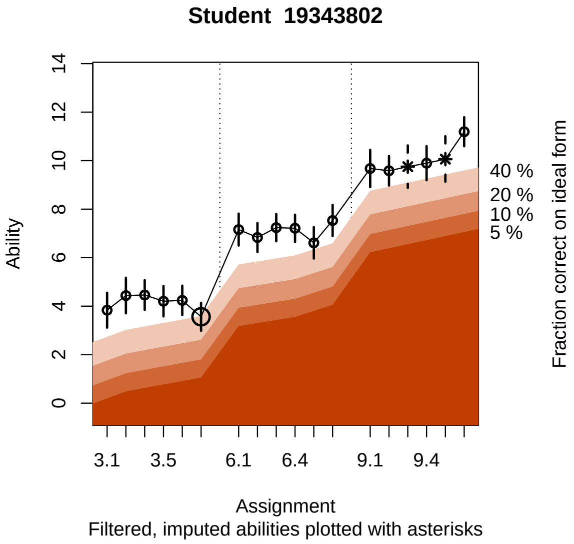

**Figure 9: Filtered Progress Map**

The filter automatically imputes ability estimates for missing assignments. The student shown in Figure 9 is missing data for assignments 9.3 and 9.5 (plotted with asterisks). The filter imputes abilities based on the previous scores. Note that the standard error grows larger for the imputed values (dashed error bars) but shrinks down when assignment data are again available. Adding thermometer plots across the top showing the completion percentage of each assignment would improve the utility of this display.

## 6. LIMITATIONS AND IMPROVEMENTS

Building rigorous psychometric models for homework is problematic because homework items are seldom selected with an eye to building a true vertical scale. By assuming that the items are assigned according to a stationary scale, we gain consistency in interpreting estimated student abilities. This allows a growth model to be used to smooth the estimates over time. The additional assumption that growth is similar between any two sections allows the ability estimates to be placed on a progress scale, which has some of the same visual appeal as a true vertical scale.

Stationarity is very much an assumption of convenience. This is both a strength, in that it allows analysis to proceed without a true vertical scale, and a weakness, in that without verification it adds an unknown bias to the ability estimates. Consequently, we only recommend the use of unverified stationary scales for low-stakes purposes, such as student ability tracking by student or instructor. High-stakes uses of the stationary scale would require verification of the stationarity assumption.

A key limitation is that the verification of the stationarity assumption requires the same kind of anchor item design as building a true vertical scale. This is part of a fundamental model identification issue: if the scale at each time slice is not identified, then neither is $\delta_t$, the average growth between time slices [1].

The stationary scale supports a variety of techniques, such as the Kalman filter, for smoothing ability estimates. We imagine that instructors will find smoothed graphs (e.g., Figure 9) more useful than unsmoothed graphs (e.g., Figure 8). A logical next step would be to evaluate Progress Map usability with instructors.

This paper modeled student ability with a single continuous variable, but stationarity generalizes in a straightforward way to the multidimensional case. The Kalman filter works as well for multidimensional normal models of proficiency. Other models, for example replacing the model of Figure 1 with a dynamic Bayesian network [5], fit into the general framework. The Kalman filter is no longer appropriate for smoothing, but the particle filter is adaptable to a wide variety of representations.

Further work is required in estimating the parameters of the growth model. The scale identification problem is insurmountable without the stationarity assumption (or a true vertical scale). But even under stationarity, estimating the growth model variance of the innovations $\omega^2\Delta t$ can be tricky. Although the Brownian motion model implies that the population variance should increase over time, that was not the case for our data set. This may be due to student attrition; the number of students actively completing assignments dropped from approximately 550 to 450 over the semester, and it is likely that the drop-outs were predominantly lower-ability students.

## 7. ACKNOWLEDGMENTS

We thank Tom McTavish, Quinn Lathrop, Simcha Knif, Brett van de Sande and Kurt VanLehn for feedback and helpful suggestions.

All opinions expressed in this paper are those of the authors and do not represent official positions of Florida State University or Pearson Education.

## 8. REFERENCES

[1] R. Almond, U. Tokac, and S. Al Otaiba. Using POMDPs to forecast kindergarten students reading comprehension. In J. M. Agosta, A. Nicholson, and M. J. Flores, editors, *The 9th Bayesian Modelling Application Workshop at UAI 2012*, Catalina Island, CA, August 2012.

[2] R. G. Almond. Cognitive modeling to represent growth (learning) using Markov decision processes. *Technology, Instruction, Cognition and Learning (TICL)*, 5:313–324, 2007.

[3] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[4] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and control.* Holden-Day, 1976.

[5] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computer Intelligence*, 5:142–150, 1989.

[6] H. Doran, D. Bates, P. Bliese, and M. Dowling. Estimating the multilevel rasch model: With the lme4 package. *Journal of Statistical Software*, 20(2):1–18, 2007.

[7] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

[8] J.-C. Falmagne. A latent trait model via a stochastic learning theory for a knowledge space. *Psychometrika*, 54:283–303, 1989.

[9] H. Huynh and C. Scheider. Vertically moderated standards: background, assumptions, and practices. *Applied Measurement in Education*, 18(1):99–133, 2005.

[10] R. E. Kalman and R. S. Bucy. New results in linear prediction and filtering theory. *Transactions ASME, Series D, J. Basic Eng.*, 83:95–107, 1961.

[11] M. J. Kolen and R. L. Brennan. *Test equating, scaling, and linking: Methods and practices.* Springer, 2004.

[12] R. J. Mislevey and R. Zwick. Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement*, 49(2):148–155, 2012.

[13] R. Patz. *Vertical scaling in standards-based educational assessment and accountability systems.* Council of Chief State School Officers, 2007.

[14] M. D. Reckase, T. A. Ackerman, and J. E. Carlson. Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25(3):193–203, 1988.

[15] L. Wise and M. Alt. *Assessing vertical alignment.* Council of Chief State School Officers, 2005.