Master Thesis Working Document

What can the parameters of IRT based learner models tell us?

Lieuwe Rekker

October 30, 2014

1. Introduction

[D: Provide short motivation for ITS's first.] Today more and more intelligent tutor systems (ITS) are being used, both in research settings and in the world at large. An ITS is a computer program used by students to learn about different subjects and generally entails that students solve problems within the system. Data from these systems have been used to build models of how students learn.[D: Most learner models do that indeed, but some restrict themselves to the probability of a correct answer... We don't like those, but still.] These learner models in turn have been used in ITSs to estimates the level of the student and adjusts the problems it presents to the student accordingly.[D: Reformulate in terms of possible usuage of student models.] In research these learner models are generally evaluated by looking at some measure of how the performance predicted by the model fits the real data.[D: Will need some cites. And, I would cute the word real.] In this research a set of learner models based on item response theory (IRT) will be discussed and it will be examined when the parameters that are fit in these models are stable.[D: Why? motivate please]

In psychology and more notably psychometrics, how students perform on problems has long been a field of research. The activity in this field was hastened by the necessity for standardized testing. This has eventually led to the development of item response theory (IRT) (for a good overview see [Ham91]). In the models stemming from IRT, students are characterized by a skill level and problems are characterized by factors such as difficulty and discrimination. Through decades of research and practical use (especially for standardized tests) IRT has gained a solid theoretical and experimental basis, which makes it a logical basis for use in ITSs in estimating the level of students. Problematic though in the application of this theory to ITS data is that learning is not taken into account.[D: Alternative, ...is that IRT assumes the compentence of the students to be constant for the duration of the test. In ITS, which are designed to help students learn, this can obviously not be assumed.] Within a test learning is not much of an issue, but in an ITS data is collected over longer time spans and learning is actually meant to occur. [D: from here onwards the structure is a bit illogical. First cluster all information you want to give about IRT by itself first, then mention why you want to extend, then mention the how, then the problems that one can encounter by doing so, that a comprehensive evaluation and comparison of these models has not yet been made - and then state what your research will be, (and maybe mention the use of your research for educators and researchers)] Therefore the adaptations of IRT models used on this kind of data incorporates a learning rate.

In IRT there are established methods to see if obtained parameters are significant as well as methods that check whether its assumptions are met. This gives confidence that found parameters for students and problems are meaningful and can be used to make statements about student skill and item difficulty. In the case of the models used for ITSs, generally only model fit compared to the data (in various forms, most notably one-step-ahead prediction accuracy) is taken into account. There are some exceptions where concerns are raised on the plausibility of parameters [GBH11], [Bec07] and accuracy of parameters [YPK11]. Not looking further into the accuracy of these parameter estimates is at the least a missed opportunity (e.g. knowing how easy certain skills are to learn, how difficult certain items are, or how fast a particular student learns can be valuable information, even if simply to improve the ITS) and might already be used wrongly in some cases (for example making statements on the difficulty of questions, without knowing whether the fitted values can be trusted).

This research will look at the influence of how certain factors influence the accuracy of parameters estimation. When applying IRT to test results, generally no external measure is used to check the found difficulty and skill parameters against, rather internal validity is used to see if the assumptions of IRT hold [Ham91]. For example, when estimating the difficulty of an item, it should not matter whether the results of low or high ability students are used to estimate the difficulty. If this is not the case, this indicates that some assumptions are not met and makes the parameter values found of questionable value. This research will take a similar approach in looking at the found parameters. The research will be split into two parts. In the first part generated data is used, so that the real parameter values are known and effects of different factors can be estimated. In the second part real data will be cut into pieces so that variance of parameters can be determined. In combination with the first part of the research it is hoped that conclusions can be drawn as to whether the parameter values found are useful and whether the data-fit measures normally looked at provide some clue to this usefulness.

2. Models

2.1. IRT Models

For every question asked an IRT model uses the student s and the item i as input and return a probability P that the question is answered correctly. The function at the basis of providing this probability is the logistic ogive function (see formula 1). Each student has exactly one parameter associated with them which represents the ability of the student and is here named skill and indicated by θ. This parameter is generally the parameter of highest interest in tests as it can be used to order the level of skill of the different students. The number of parameters associated with items and how all these are combined into x differs per particular model and is discussed in the following paragraphs.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

2.1.1. 1PL or Rasch Model

The 1PL model, also known as the Rasch model has one parameter (b) per item (hence the name), which stands for the difficulty of the item. The entire formula for a question's probability of a correct answer is then: $P(s,i) = \sigma(\theta_s - b_i)$. This means that when the skill of the student and the difficulty of the item are on a par, the student has a probability of .5 to answer the question correctly. Note that there exists an indeterminacy issue with this model: one can keep the same model while changing the parameter values by increasing or decreasing all θ and b by the same amount. This problem is generally solved by setting the average θ to 0.

2.1.2. 2PL Model

The 2PL model expands the 1PL model with the parameter a which is called the discrimination of the item. The 2PL then looks like this: $P(s,i) = \sigma(a_i(\theta_s - b_i))$ The term discrimination comes from the fact that a high discrimination causes P to change quickly when $(\theta_s - b_i)$ is small and thus the performance of students who are close in skill can be more easily distinguished. The flip-side here is that when $(\theta_s - b_i)$ isn't small, P will more quickly drop to 0 or rise to 1, concealing any difference between the skill levels of those students. Note that for this model not only can θ and b be increased or decreased by the same amount, but all a can be scaled up or down to will as long as all θ and b are scaled down or up respectively by the same factor. This problem is generally solved by setting the average θ to 0 and its variance to 1.

2.1.3. 3PL Model

The final IRT model discussed here is the 3PL model which adds a chance parameter c. This model takes into account that on occasion the student could answer a question correctly by taking a (educated) guess. This phenomenon is of course most prevalent in multiple choice tests where the chances of correctly guessing the answer are relatively high. The model effectively changes the lowest probability to the level of c and the space between c and 1 is rescaled accordingly leading to formula 2. This model suffers from the same identifiability issues as the 2PL model. In this thesis this model is left out of consideration, as none of the learning models are based on it and none of the data is multiple choice and will thus not be mentioned again.

$$P(s,i) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta_s - b_i)}} \tag{2}$$

2.1.4. Fitting the Models

The above paragraphs have discussed what the parameters are taken to represent and how they are used in the model. None of these parameters are directly observable (i.e. they are latent) and thus are not directly obtained from observations. Instead the observed correctness of the answers to the questions asked and the model probabilities for the correctness of those answers are used. The parameters are given those values at which the likelihood that the observed answers arise from the model is maximized. The likelihood of a single question is the probability that the correctness of the answer is seen according to the model. Thus if the question is answered correctly, the likelihood for that answer is P, while if the question is answered incorrectly the likelihood is $1-P$. By multiplying the likelihood of every datapoint the likelihood of the entire dataset is determined and we end up with formula 3.

$$L = \prod_{d \in D} P_d^{t_d}(1-P_d)^{1-t_d} \tag{3}$$

In the 1PL model x is linear in the parameters, which makes maximizing the likelihood quite straightforward: logistic regression can be directly applied to this problem. In the 2PL (and 3PL incidently, which will now really not be mentioned further) x is bilinear and thus logistic regression can not be applied directly. Instead values for student parameters are fixed (making x linear again), logistic regression is applied, the found item parameter values are then kept fixed to find the values for the student parameters. This procedure is repeated until the likelihood of the data is (almost) the same in different runs of logistic regression. For more detailed information on logistic regression and how it is used here, please refer to appendix A.

2.1.5. Assessment of fit

The issue with IRT is that the parameters which are of interest are not directly observable. It might be possible through the use of experts to get an indication of what values, but this would at least be costly and still difficult to do. In [Ham91] puts forward three types of evidence to inspect model fit:

> [J]udgements about the fit of the model to the test data be based on three types of evidence: 1. Validity of the assumptions of the model for the test data 2. Extend to which the expected properties of the model (e.g., invariance of item and ability parameters) are obtained 3. Accuracy of model predictions using real and, if appropriate, simulated test data.[p.55]

The third point is most straightforward: if the predictions by the model for questions that were not used in fitting are inaccurate, the model is probably not a good fit for the data. The most important assumption the first point refers to is that of unidimensionality: the skill represented by θ should be the only skill of importance in answering the items. A good example is when some questions on a math question use difficult wording, making a high skill in math insufficient for correctly answering this item. The second point and especially the mentioned example of invariance is of major importance for IRT. In Hambleton's own words: "The property of invariance of item and ability parameters is the cornerstone of IRT". Invariance means that the values obtained for items are the same whether they are fitted on the data obtained from one group of students or another group of students and the same for the parameter values of students.

——

All models used in this research employ an important extension: it incorporates learning by students. Skill, θ is split up into an initial part and a learning rate, so that each time a question is answered the skill of the student increases. In the ITS there are many different knowledge components and thus the parameters are fit per knowledge component. Depending on the data-set used an item can have multiple components associated with them, creating the necessity to combine multiple item response functions. ——

Say something about the fitting procedure for each seperate model.[D: what, here? or do you mean the alternate ML optimization of student and item parameters?]

[D: List the assumptions each model makes. Either in each subsection, or in a separate subsection]

2.2. Learning Factor Analysis (LFA)

LFA uses a simplified version of the IRF, but extents it by introducing a learning rate as discussed in the introduction and by allowing multiple knowledge components to be associated with a single item. The combination of KCs is made by summing the learned part of knowledge and the difficulty of the KC for every KC that is linked to the item.

$$P = \sigma(\theta_{s,0} + \sum_{c \in KC} \eta_c t_{s,c} - \beta_c) \tag{4}$$

[D: So the start competence is the same for every knowledge component/skill, but the learning speed can vary per KC/skill... that is an assumption that is explicitly rejected in the eIRT paper.]

The simplification of the model is done by dropping α (this model is called the Rasch model). The splitting of θ leads to the introduction of an initial skill θ_0 defined per student, a learning rate η defined per KC (i.e. the KC determines how fast or slow learning occurs) and a number of times that a student has seen items belonging to this particular knowledge component $t_{s,c}$. Please note that in the the original LFA β is added. It is subtracted here to maintain similarity to the original IRF and ensure uniformity with the other models used. This has no other effect than that the signs for β are reversed.

When looking at a data set where only a single knowledge component is linked to every item, β and θ_0 are not independent: we can raise both by any amount. To solve this issue, the average value of θ_0 is set to 0.

2.2.1. Combining knowledge components

Something more can be said about the way KCs are combined. In [CKJ08] Cen et al. show that in practice there is no difference in performance between an additive model (as used here) or a conjunctive model (where probabilities of individual KCs are multiplied). Nevertheless the authors already mention that this is probably the case because for most KCs $\beta < \eta_c t_{s,c}$, meaning that adding KCs does decrease the chance of answering the question correctly as would be expected. In their paper they already propose using a data set where ($\beta > \eta_c t_{s,c}$) to see if this is indeed why this way of combining KCs works well in practice.

The experiment proposed above is put to the test here. Fitting a conjunctive model is hard in practice, but generating data using one is rather straightforward. Whether a real life data set contains many questions where skills are such that $\beta < \eta_c t_{s,c}$ cannot be said at this point. Even if this is not the case though, it can be argued that the parameter values can be skewed slightly to ensure that this occurs. The rationale behind is, is that the fitted values obtained from the real data may be skewed towards $\beta < \eta_c t_{s,c}$ due to a additive model being used in the fitting process. It would then be expected though that the retrieved parameters using these values would be skewed towards $\beta < \eta_c t_{s,c}$ again.

2.3. Performance Factor Analysis (PFA)

PFA is a direct extension of LFA. In PFA separate learning rates are used for questions answered correctly and questions answered incorrectly. Additionally θ_0 is dropped. As put forward in [PCK09b] θ_0 is dropped because this extension is made mostly to make this model more useful in ITSs. Leaving out any student specific parameter makes the model more easily applicable to students not used in the fitting procedure. As noted in both [GBH11] and [YPK11], leaving out $\theta_{s,0}$ makes parameter estimates worse. Since prediction for students who were not part of the fitting procedure is not a primary concern here, a model that does include $\theta_{s,0}$ (as done in [GBH11] and [YPK11]) is used instead of PFA and will be referred to as PFA+.

$$P = \sigma(\theta_{s,0} \sum_{c \in KC} \gamma_c g_{s,c} + \rho_c f_{s,c} - \beta_c) \tag{5}$$

[D: Is there a minus sign missing here? same assumption as the last, but two different learning rates.]

Here γ is the learning rate of the KC for correct answers and g is the number of questions answered correctly. Consequently ρ is the learning rate of the KC for incorrect answers and f is the number of questions answered incorrectly. Just as with LFA above the sign for β was reversed compared to the original.

2.4. extended Item Response Theory (eIRT)

The extended Item Response Theory model by Roijers et al [RJF12] is the most straightforward extension to the standard IRT model.

$$P = \sigma(\alpha_c(\theta_{s,0} + \eta_s t_{s,c} - \beta_c)) \tag{6}$$

[D: Only one competence and learning speed.]

Similar to LFA θ is replaced by initial skill and a learning rate. Here the learning rate is taken per student though rather than per knowledge component as is done in LFA and PFA. With θ split up, it would seem that α obtains a slightly different meaning. For θ_0 it still has the same discriminatory function. When looking at η though, α directly impacts it as a modifier, making learning easier (¿1) or more difficult (¡1).

Actually I have no clue on the correct mathematical vernacular, so please nudge me to the right terms to make this more understandable/consice It should be noted that different parameter settings can lead to exactly the same model. E.g. all θs and βs could be increased by the same amount and the model would still be the same. To still be able to compare parameter values and variances the parameters should be normalized as follows. The average of $\theta s, 0$ will be set to zero as to resolve the dependency with β values. The standard deviation of θ will be set to 1 as to fix the dependency between α and the other parameters.

2.4.1. Adapting eIRT

eIRT as defined by Roijers et al does not incorporate multiple skill steps. In order to be trained on multi-skill data and to be similar to the other models 7 will be used as a multi-skill extension of the eIRF. To distinguish this version from the original eIRT, this extended version will be denoted as seIRT. A notable difference here is that $\theta_{s,0}$ is divided by the number of KCs involved. This is because $\theta_{s,0}$ should only be added once just as in LFA/PFA, but nevertheless it should be modified by the corresponding α_cs as well.

$$P = \sigma(\sum_{c \in KC} \alpha_c(\frac{\theta_{s,0}}{|KC|} + \eta_s t_{s,c} - \beta_c)) \tag{7}$$

Please note that in the multidimensional case the dependency between $\theta s, 0$ and β is most likely no longer there. The dependency between α and the other parameters still exists though and should still be resolved by fixing the standard deviation of $\theta_{s,0}$ to 1.

2.5. Combined Model

The three models introduced above can all be encompassed by a more complex model.

$$P = \sigma(\sum_{c \in KC} \frac{\alpha_c \theta_{s,0}}{|KC|} + \eta_s \gamma_c g_{s,c} + \eta_s \rho_c f_{s,c} - \beta_c) \tag{8}$$

LFA can be obtained from this model by taking $\alpha = 1$, $\eta = 1$ and $\gamma = \rho$. PFA can be obtained from this model by taking $\alpha = 1$, $\eta = 1$ and $\theta_0 = 0$ (minus the last one for PFA+). The adapted eIRF can be obtained by taking $\gamma = \rho = \alpha$ and realizing that β already incorporates α.

3. Related work

[D: Right now, this section is a laundry list: it could be improved by adding some more structure.]

In [Bec07] Beck goes beyond investigating the accuracy of a model (knowledge tracing in this case) and also looks at the parameter values. The authors prime reason for concern lies in identifiability: the fact that widely differing parameter settings can lead to almost identical model outcomes. Although this paper does concern itself with the 'plausability' of parameter values it only goes so far as to nudge the parameter to values deemed plausible rather than asking the more fundamental question of whether the parameter values are useful at all.

Learning factor analysis (LFA) [CKJ06] is a cognitive model of how students learn. It is a Rasch model[D: is one of the IRT variants... (you have not mentioned that yet)] extended with a learning rate to model learning over time. In [CKJ06] LFA is used in the context of a greater cognitive model of learning and skill that includes what knowledge components are linked to what items. The quality of fit (measured as the log likelihood of the entire data set given the model plus some penalty for the number of parameters) is used as a measure of quality of a specific linkage of knowledge components to items.

Performance factor analysis (PFA) is a further extension to LFA. It is introduced in [PCK09b] as an alternative to knowledge tracing and focuses more on correctly estimating whether a student has mastered a particular knowledge component. The difference with LFA is that learning rate is dependant on whether a question is answered correctly or not. Furthermore initial student knowledge is dropped. The foremost reason for doing so is to make this model work for students of whom no data had been taken into account when fitting the model.

In [GBH11] Gong et al. also made a comparison between various knowledge tracing approaches and PFA. Whether PFA or KT performs better remains inconclusive. Upon inspecting parameter values they found that many learning rates were negative, which seemed implausible in real life. They noted that upon placing a lower bound of 0 on the learning rate, performance improved. Additionally the authors used a pretest and correlated the performance on this test to the initial knowledge parameter from the model. In this set-up an adapted version of PFA (the same version that will be used in this research: the original doesn't use initial student knowledge) showed the highest correlation).

In [YPK11] Yudelson et al. show some particular factors that can negatively influence the quality of PFA models. One of the factors looked at here is model complexity: on the one hand this is done by using a more finegrained set of KCs and on the other hand by adding another parameter to the PFA model. In evaluating their results the authors did not only look at accuracy measures, but also inspected values for specific parameters. *how did they get their standard deviations for the parameters? They don't seem to have split their data?* In inspecting learning parameters they also noted how the learning parameter for wrongly answered questions are often negative when initial knowledge is not included in the model. They concluded that PFA not so much models student learning in this case, but rather performs some kind of 'error tracking' in order to produce good estimates of skill despite not having any additional information on the student. They therefore prefer an adaptation of PFA that includes initial student knowledge.

In [RJF12] Roijers et al. extend the 2PL IRT function to include a learning rate. This new version is called extended IRT (eIRT). There are two important differences in how this model has been extended and how LFA and PFA came to be. First LFA and PFA use a multidimensional model, which means that multiple skills can be associated with a single problem. Moreover in eIRT the learning rate depends on students, while in LFA and PFA the learning rates depend on the skills. Roijers et al. perform experiments using generated data, that shows that parameters can often be retrieved from the data. Additionally they have a real life data-set based on 14 students from three different groups. With the small amount of data in their real life set, they conclude based on their previous experiment that only rule difficulty and student initial knowledge can be determined with enough accuracy. After training the model on the data the fitted initial knowledge of the three different groups confirmed their hypothesis of the ordering of average initial knowledge per group. Also the ordering of difficulty of the rules involved is indistinguishable to rankings made by experts. The authors thus showed that the difficulty and initial knowledge parameters obtained reflect the ordering of these in real life.

4. Research Question

The main question looked into in this thesis is ”When are parameter values found in fitting IRT based models distinguishable?”. The interpretation of this question is specified in this section and some subquestions are defined which provide structure for the conducted experiments and the discussion of the results. The questions are each formulated in a subsection and referred to throughout the rest of this thesis.

Distinguishable here refers to the need of knowing if the values of parameters are distinct rather than knowing how reliable any particular value is. I.e. the important question would be 'If the value for Tim's initial knowledge is higher than John's is this actually correct?' Thus instead of looking at how reliable particular values obtained in fitting the model are, it is looked at how reliable the ordering of the different values are.

There are different levels on which distinguishable can be looked at. In the question above it is on the level of two specific parameters. The default level on which this question will be answered is on the level of parameter type: how well are all the parameters of a particular type distinguishable from each other (macro level). Nevertheless some energy is put into exploring this question on the micro level: i.e. can we say something about how distinguishable a specific parameter is from other parameters?

When on the one hand refers to the fact that distinguishable will depend on some factors, such as how much and what data is used in fitting the model. When also refers to finding conditions that indicate how distinguishable the parameters are. In later sections it is seen that finding the distinguishability of parameters is complex, time consuming and restricted to subsets of the data. Thus finding other, more easily obtained measures that indicate distinguishability would be favorable.

4.1. What is the influence of the amount of data on distinguishability

It is expected that simply increasing the amount of data while holding the number of parameters constant will improve the estimation of those parameters and thus increase distinguishability.

4.2. How does distinguishability differ between domains?

There are some quite different ITSs that provide data in a form that is suitable for the IRT based models under investigation. The structure of the data between these ITSs and even between different subjects within the same ITS can be quite different. Additionally, learning is a complex matter for which many differences between classes using the same ITS can already impact how well the model works. Therefor different datasets are looked at to represent different domains.

4.3. How can we easily know that parameter values are distinguishable?

This subquestion refers to the second interpretation of 'when' in the main research question: for some measures that are more easily obtained than the distinguishability of parameters it will be checked how well they relate to the distinguishability of parameters, both on the macro level and the micro level.

4.4. To what extent is distinguishability attributable to stochasticity of the model?

The models used are stochastic in nature. Even when the probability of a correct answer is 90%, an incorrect answer is still expected in 10% of the cases. This means that even if data would be noiselessly generated by the model, some variation in data and thus some variation in found parameters will be observed. It is expected that noise in the domain and mismatch between model and reality will add additional variance to this. To separate the two and to see if maybe this variance (which is easier to obtain) correlates with the total variance, it is studied as well.

5. Method

5.1. Model Performance

The most common performance measure used in the context of ITS learning models is some measure of accuracy of model predictions on next-item student performance. Although this research looks into the values of model parameters and is less concerned with other measures of performance an accuracy measure is still used. The main rationale behind this relates to subquestion 4 as the accuracy of a model may hold some relationship to how well the parameters are matched. This would be especially useful as accuracy measures are relatively easy to obtain and are often already looked at in ITSs.

The specific accuracy measure used for this research is A' (pronounced a-prime) [DB12]. For this measure two items are represented to the model, one which was answered correctly and one which was answered incorrectly. The model is used to determine which is which. The advantage of this method is that "values of A' are statistically comparable across models and data sets" [DB12]. *there is a problem in using normal A' due to the dependencies of answers by the same student, there is a paper that deals with this, which should be incorporated*

Log likelihood of the data given a model is another measure that plays a large role in this specific context as this is what is maximized in fitting the model. Although the log likelihood value might not be very informative, it does give an indication of how well the model fits the data-set and in combination with the measured accuracy one can get an idea to what extent over-fitting occurs. To be able to compare this value over different runs it will be normalized by dividing by the number of data points.

currently this part is out of the question. I might bring up this discussion again: Additionally a method that from IRT [Ham91] will be used. After initially fitting the data will be split in two along skills, such that one set contains all the easiest skills and one will contain all the most difficult skills. Then the fitting will be done again on both new data-sets. If the assumptions of the model were correct the parameters for students found should be roughly the same from both sets. This procedure will be repeated, but then making a split on student initial skill and looking at the skill parameters.

Although the primary way in which model parameters will be examined is described above, alternative methods will be used where applicable. Some data-sets contain additional information that can be used to more directly draw conclusions on whether found parameter values are meaningful. Two examples from the related research section are [RJF12] where expert opinions and an indication of what groups have higher initial knowledge were used and [GBH11] where a pre-test was used as an indicator of initial knowledge.

5.2. Inherent Variance

Inherent variance is the variance discussed in subquestion 1. In IRT the concept of a test information function is used for this. In short this function is defined for a set of items and (indirectly) gives the variance for any student knowledge parameter evaluated on these items. Baker (2001) describes that the found variance is the variance that would be observed on fitting a model where a student would answer this set of items over and over again, while every time forgetting he has seen these items. An equivalent can also be made to estimate the variance of item parameters or both parameter sets at the same time. The variance of these information functions is only dependent on the value of all the parameters involved and of the items asked to the student. The labels of the items are thus irrelevant for the information function.

Such a function has the drawback that this variance is only achieved in the limit. I.e. the actual variance approaches this variance closer and closer as the number of observations go up. Some of the datasets looked at here are rather small though. A simulated approach will thus be used to determine the inherent variance of the model. The information function will still be used though and compared to the empirically found inherent variance as the information function is relatively easy to obtain and could be accurate enough and play a role in answering subquestion 4.

In the simulated approach, first the parameters are determined by fitting the model on the data (here the labels do have their influence on the result). The found parameters are then used to stochastically generate new labels for the data. This means that if the model predicts a .2 probability that the question is answered correctly, 20% of the time the label will be 1 and 80% of the time the label will be 0. In the PFA model this value will also influence further probabilities due to the different learning rates for correct and incorrect answers to questions. Multiple sets of labels are generated for the data in this fashion, after which the same model is fitted again on these sets. The inherent variance of every parameters is then estimated by calculating the variance of that parameter over the different trained models.

5.3. Observed variance

In order to get an idea of the variance of parameters in reality, data is split into multiple parts. The same model is fitted on each part of the data (from now on called a split) and the variance for each parameter can then be calculated over the different models.

The focus here lies on the KC parameters. Also the splitting of the data should retain the way the data might be seen by an ITS in real life. Therefor the data is split on the basis of students: every split contains the data of some of the students and every student in a split does not occur in another split. The same split is done on the test set such that the model trained on the data of specific students is also tested on the test set of those particular students.

To investigate subquestion 2, different splits are made, with a decreasing number of students per split. It should be noted that this means that for the larger splits, a lower number of parameter values are present and that thus the second order variance will be higher.

5.4. Representation of variance

The variances discussed in the previous sections are not easily interpretable as it is not directly clear whether a variance is large or small for a particular parameter. To account for this problem the following reliability measure for every parameter value will be used:

$$R(\beta_n) = 1 - \frac{\sigma^2(\beta_n)}{\sigma^2(\beta_n) + \sigma^2(\beta)} \tag{9}$$

Where β_n is the β parameter of the n-th KC, $\sigma^2(\beta_n)$ is its variance and $\sigma^2(\beta)$ is the variance over the β parameter of all KCs in the model. The reliability can be calculated for both the inherent and observed variance and can be calculated for any parameter, although β was taken as an example here. Two issues should be noted here. First

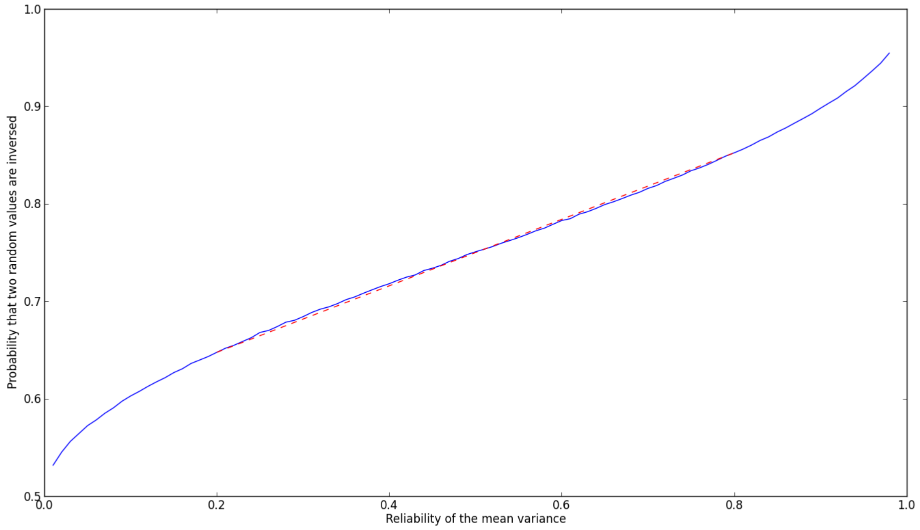


Figure 1: Relation between reliability and percentage of correctly distinguished parameters

a normal distribution of the parameters is assumed here and when calculating from different reliabilities of different parameters the harmonic mean should be used.

When considering the consistency of parameters, a good way to look at them is distinguishability. For example, if we were to look at every possible pair of parameters, in how many cases would the one that is seemingly largest, actually be the largest? The reliability over the average variance of a parameter type is directly related to this measure (see figure 1)

5.5. Domains

In generating the data there are many other factors that might play a role in the performance of the models. Among those factors are the values that the parameters will be given, the distribution of knowledge components over the items, the ratio of students to items etc. As indicated in section 4 these factors are not explored methodically and extensively, but rather a few different datasets are used to represent some of the variation naturally found in this kind of data. The different datasets and some of the characteristics are described in this section.

5.5.1. Bridge to Algebra

The first data-set is one of the datasets used in the 2010 KDD cup on education data mining. The data is from Carnegie Learnings' Cognitive Tutor Software meant for high school-students aged 15-18. [RLK^{+}08] provides an overview of some of the features of this program. The data was obtained from the Bridge to Algebra course during the school year 2006-2007 [SNMR$^+$10b].

5.5.2. Algebra I

This dataset was also provided in the 2010 KDD cup. The data is also obtained from Carnegie Learnings' Cognitive Tutor Software, but from the Algebra I course during the school year 2005-2006 [SNMR$^+$10a]

5.6. Assistment

This dataset is quite different from the other two as it is taken from a web-based ITS named Assistment whose details and development story can be found in [**?**]. The data is from 12 to 14 year old students and was used in [GBH11] which was discussed in section 3. The data does not only contain data from usage of the ITS but also data from a pre-test.

5.7. Data cleaning

Splitting the data may exacerbate some of the issues that can be encountered in fitting the model. One example is when all questions associated with a student or a KC are answered correctly or incorrectly. This makes the fitting algorithm want to assign infinite values to parameters. Another problem is when for a KC there is such a limited number of questions answered that learning rates cannot be estimated.

To prevent these issues the following steps are taken. On the whole dataset, students that answer all questions correctly or incorrectly are removed. In every split KC's for which every question is answered correctly or incorrectly is not taken into account for that split. Additionally if for a KC there is less than two questions answered correctly or incorrectly that KC is not taken into account for that split. This means that this KC is removed from the items and any item that no longer contains a KC is dropped from the data. It is assumed that this does not cause students to now have answered all questions correctly or incorrectly, but this assumption will be checked during the experiments to be certain.

5.8. KC Experiment

5.8.1. The Experiment

In detail the experiment for KC parameters consists of the following steps per model:

1. Clean the data by removing students according to section 5.7
2. Split the data in a increasing number of parts. Then for each part:
   a) Clean the data by removing KCs according to section 5.7
   b) Train the model and determine the variance for every parameter type

c) Determine performance on the test-set for that split

d) Determine the inherent variance through the Fisher matrix and through simulation

e) Normalize the found variances with the variances per parameter type found above

3. Determine the variance per parameter over all the splits

4. Normalize the found variances with the average of variances per parameter type found above

5.8.2. Representing the Results

Dit is een voorstel voor het representeren van de resultaten. Feedback, ideeen of suggesties worden erg gewaardeerd. For every split into a number of parts and every model the following will be represented:

1. The average reliability for each parametertype (both inherent and observed *preferably the reliability of the difference between the two as well, even though that has not worked too well in the testrun I have shown, although I'm less than 90% sure I did not make a mistake*)

2. The average accuracy and normalized log likelihood

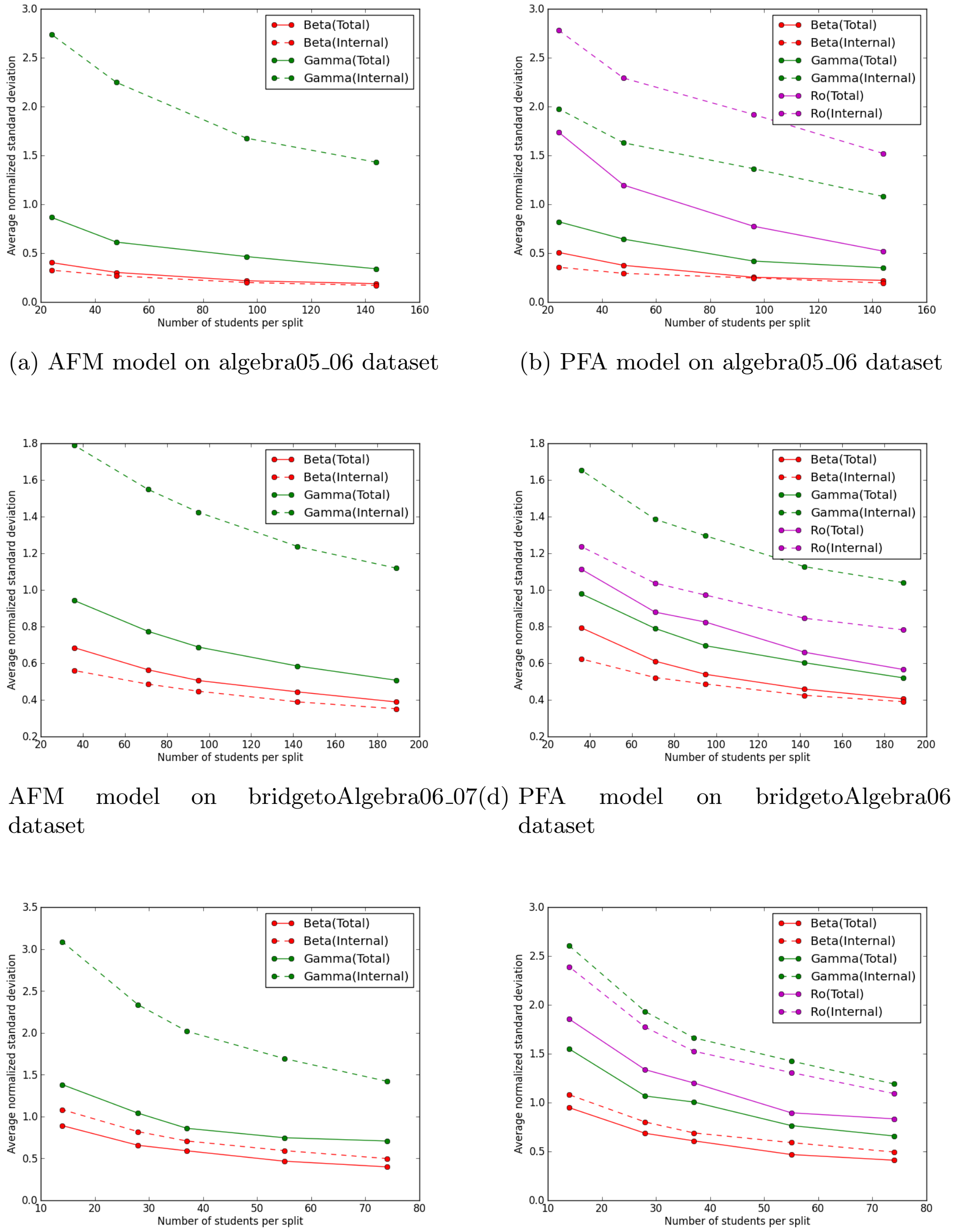
3. Histograms that display the reliabilities of the different parameters. Outliers can now be easily spotted.

I'd also like to see to what extent inherent and total variance are related over the different splittings and models, as well as relations with accuracy and normalized log likelihood. correlation or spearmans R are two options I'm considering

6. results

Kendall A few observations on the basis of the figures below. Foremost is the observation that the inherent variance is sometimes higher than the actual observed variance. A possible conclusion might be that reality is not quite as stochastic as the model: you either know a knowledge component or you don't and there is far less of a chance of answering correctly when you don't know an answer or vice versa. The question what part of the variance is due to internal variance is not useful as sometimes the variance is above and sometimes below the variance that is observed. This internal variance might still be useful though as it is rather correlated to the total variance, especially at lower values. (see the figure below)

As expected the variance is less as more data is available although the variance seems to decrease less as data is increased. Also quite obvious is that the learning parameters have far higher variances than the beta parameter. Interestingly adding an additional parameter does not substantially change the variance of the beta parameter.



(a) AFM model on algebra05_06 dataset

(b) PFA model on algebra05_06 dataset

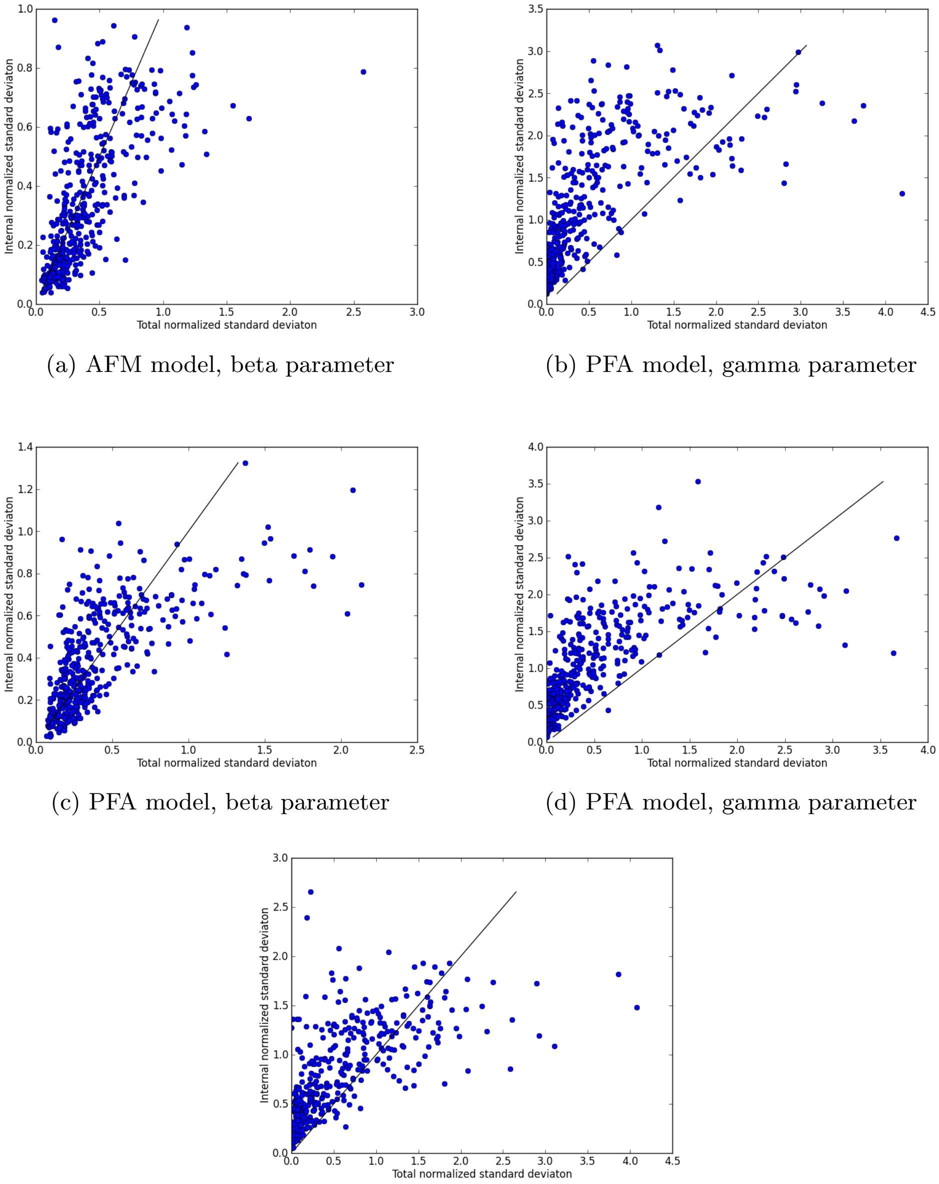
(c) AFM model on bridgetoAlgebra06_07 dataset

(d) PFA model on bridgetoAlgebra06_07 dataset

(e) AFM model on assistmentGong dataset

(f) PFA model on assistmentGong dataset

Figure 2



(a) AFM model, beta parameter

(b) PFA model, gamma parameter

(c) PFA model, beta parameter

(d) PFA model, gamma parameter

(e) PFA model, ro parameter

Figure 3: Standard deviations on the values of individual parameters on the bridgetoAlgebra06_07 dataset

7. Things to add and change

- *Describe terms and use those consistently!*
- *implement and describe the adapted A' approach*
- *related work is a bit out of place as it is not very intelligible without information given later. Maybe move it to the end, or rather integrate it with other stuff?*
- *Describe the datasets and add some extra info on what the datasets look like?*
- *Investigate some of the outliers?*

References

[Bec07] J.E. Beck. Difficulties in inferring student knowledge from observations (and why you should care). In *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*, pages 21–30, 2007.

[CKJ06] Hao Cen, Kenneth Koedinger, and Brian Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems*, pages 164–175. Springer, 2006.

[CKJ08] Hao Cen, Kenneth Koedinger, and Brian Junker. Comparing two irt models for conjunctive skills. In *Intelligent Tutoring Systems*, pages 796–798. Springer, 2008.

[DB12] M.C. Desmarais and R.S.J. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, pages 1–30, 2012.

[GB11] Y. Gong and J. Beck. Items, skills, and transfer models: which really matters for student modeling. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 81–90, 2011.

[GBH11] Y. Gong, J.E. Beck, and N.T. Heffernan. How to construct more accurate student models: comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education*, 21(1):27–46, 2011.

[Ham91] Ronald K Hambleton. *Fundamentals of item response theory*. Sage Publications, Incorporated, 1991.

[KMS] K.R. Koedinger, E.A. McLaughlin, and J.C. Stamper. Automated student model improvement.

[PCK09a] P.I. Pavlik, H. Cen, and K.R. Koedinger. Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 121–130, 2009.

[PCK09b] P.I. Pavlik, H. Cen, and K.R. Koedinger. Performance factors analysis–a new alternative to knowledge tracing. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 531–538. IOS Press, 2009.

[RJF12] Diederik M. Roijers, Johan Jeuring, and Ad Feelders. Probability estimation and a competence model for rule based e-tutoring systems. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 255–258, New York, NY, USA, 2012. ACM.

[RLK+08] Steven Ritter, Carnegie Learning, Kenneth R Koedinger, William Hadley, Albert T Corbett, and Marsha Lilly. Classroom integration of intelligent tutoring systems for algebra and geometry. In G. Blume and K. Heid, editors, *Research on Technology and the Teaching and Learning of Mathematics: Vol. 2 Cases and Perspectives.* Charlotte, NC: IAP., 2008.

[SNMR+10a] J. Stamper, A. Niculescu-Mizil, S. Ritter, G.J. Gordon, and K.R Koedinger. Algebra i 2005-2006. development data set from kdd cup 2010 educational data mining challenge. http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp, 2010.

[SNMR+10b] J. Stamper, A. Niculescu-Mizil, S. Ritter, G.J. Gordon, and K.R Koedinger. Bridge to algebra 2006-2007. development data set from kdd cup 2010 educational data mining challenge. http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp, 2010.

[YPK11] Michael Yudelson, Philip Pavlik, and Kenneth Koedinger. User modeling–a notoriously black art. *User Modeling, Adaption and Personalization*, pages 317–328, 2011.

A. Implementation and Mathematical argumentation

The AFM and PFA model are relatively straightforward in their data representation and implementation. For these models x in formula 1 is linear in the parameters, which means that standard logistic regression can be applied. In this appendix first the way the data is represented is described followed by a proof that logistic regression indeed finds the parameter values where the likelihood of the data is highest.

A.1. Data Representation

For logistic regression the data is represented in a matrix Φ such that Φw is equal to a vector of each value of x in formula 1, where w is a column vector of the parameter values. In this matrix the rows represent data points while the columns represent what the parameters should be multiplied with. The dimensions of the matrix are thus equal to the number of data points by the number of parameters.

In the formulas for AFM (formula 4) and PFA (formula 5) x consists of a sum where every part contains exactly one parameter. This makes construction of Φ straightforward: in each row (thus for every data point) a 1 is placed for every present parameter that stands isolated (this goes for θ and β) and for the others (η,γ and ρ) the right value for that data point is inserted (a non-negative integer). Any parameter not used for that specific datapoint will have a value of 0.

A.2. Workings of Logistic Regression

Logistic regression estimates the values of the parameters that maximize the likelihood of the data given the model. The likelihood of the data is equal to $\prod_{d \in D} P_d^{t_d}(1-P_d)^{1-t_d}$

B. Glossary

Making a start with using terms consistently, plus I generally feel that a glossary would have helped me greatly in understanding papers etc.

Item A problem (step) in the ITS to which a single answer can be given

Knowledge Component A skill, piece of knowledge etc. that is associated with one or more items and in which students can have a level of competence

Question An instance of an item

Skill Level of an instance of a knowledge component for a particular student