Using a Hierarchical Model to Get the Best of Both Worlds: Good Prediction and Good Explanation

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15206
koedinger@cmu.edu

Lu Sun
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15206
ls1@andrew.cmu.edu

Elizabeth A. McLaughlin
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15206
mimim@cs.cmu.edu

ABSTRACT

Understanding how learning transfers from one task to another is a critical topic in learning science. In this paper, we investigate the impact of the scope and granularity of learning transfer by comparing three models across multiple data sets. Prior work demonstrated the value of component models of learning transfer that group items into knowledge components. Within the component models, that work left open whether difficulty (variation in performance over tasks) is better modeled by knowledge components (the strong model) or by items (the "weak" model). The strong component model is theoretically desirable because it provides a single explanation for both difficulty and transfer. However, we find that the weak component model better predicts student performance across six data sets. While this weak model predicts better, it is hard to interpret because an explanatory parameter that represents latent knowledge difficulty of student performance is absent. To maintain explanatory power without sacrificing prediction, we propose a new alternative that uses a hierarchical mixed effect regression model where item difficulty is pooled within component difficulty. Experimental results, across six data sets, show that the predictions of the hierarchical model are better than the strong model and as good as the weak model, while also producing theoretical useful explanatory parameter values for knowledge components.

Keywords

Transfer, hierarchical mixed effects models, student modeling, knowledge component modeling

1. INTRODUCTION

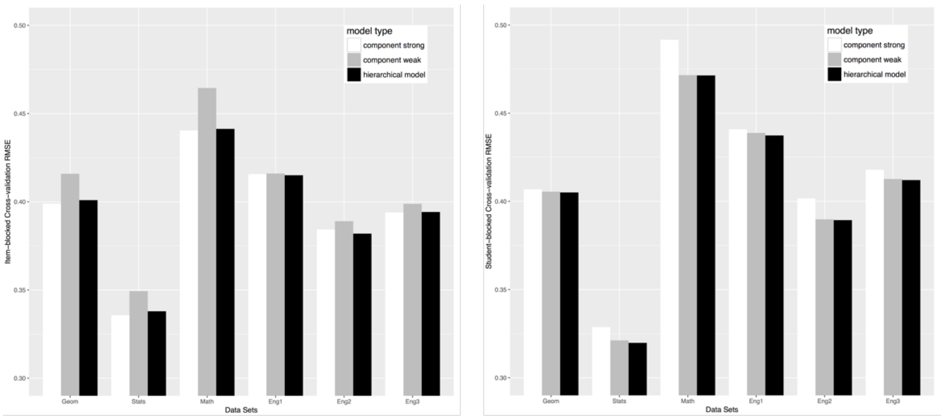
Transfer of learning, the application of knowledge acquired in one situation to other new, relevant learning situations, is an age-old fundamental problem in human cognition and education [1, 8]. A central question is determining the loci of transfer at a grain size of analysis that is fine grained to make accurate and useful predictions yet broad or simple enough to provide explanatory insight that advances science or application. In modeling transfer, [4] contrast two statistical models of the faculty theory of transfer with two statistical models of an alternative component theory of transfer using multiple datasets. That work provided a convincing case for a component theory of transfer over a faculty theory of transfer, it also raised a new question. To be effective, statistical models of learning transfer control both for general student proficiency and variations in the difficulty of tasks. When contrasting statistical models of the component theory, the results were mixed as to whether task difficulty is better modeled by items or by components.

It is worth stating more precise definitions for key terms: item, knowledge component, strong and weak component models. For our purposes, *items* are tasks that appear as questions or steps in problems where student responses are evaluated as correct or not. An example item is to find the area of a circle given its radius is 10. A *knowledge component*(KC) is defined as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks"[3]. For example, analysis of student correctness data on related geometry tasks leads to inferences about differences in generality of KCs related to trapezoid area versus circle area. Whereas few differences in difficulty across trapezoid items suggests a single general KC: "Use the trapezoid area formula to find any unknown value in the formula", difficulty differences across circle items suggests two KCs that separate items into two groups depending on whether the area is unknown (easier) or the radius is unknown (substantially harder)[5].

A *strong component model*[4] is one in which item difficulty and learning transfer are *both* modeled using KCs. A *weak component model* uses KCs only to model transfer but uses items to model item difficulty. It is "weaker" because it provides a less coherent and less parsimonious theory by having separate explanations for difficulty and transfer rather than a unified explanation as in the strong model. Despite this explanatory disadvantage, the weak model was found to produce better predictions than the strong model in some cases[4]. In particular, it produced better predictions than the strong model when generalizing to unseen students. Given the theoretically desirability of the strong model, we set a goal to develop a statistical model that would maintain the theoretical benefits of the strong model without any sacrifice to prediction. To address this goal, we used hierarchical mixed effects regression as an approach that allows the combination of estimates of both item difficulty and component difficulty.

2. RELATED WORK

Work on statistical student modeling has pursued a variety of alternatives within families of logistic regression variations, of Bayesian Knowledge Tracing variations, and of recurrent neural networks. Logistic regression variations include a few simpler alternatives, like Item Response Theory[10] and the Additive Factors Model (AFM). Recently, a family of statistical student modeling based on



a. Item-blocked cross validation

b. Student-blocked cross validation

Figure 1: Root mean squared error (RMSE) results for item-blocked (a) and student-blocked (b) cross-validation across 6 diverse datasets showing a disadvantage of the weak model in generalization across items (middle gray bars are highest in a) and a disadvantage of the strong model in generalizing across students (left white bars are highest in b). The hierarchical model provides the best prediction fit in most cases (8 of 12) and is essentially tied with the strong model in item generalization in the other cases.

recurrent neural networks has begun to emerge[7]. Deep Knowledge Tracing is highly complex and is particularly difficult to interpret, thus limiting its explanatory power and application potential, at least at this point. Among these methods, key reasons for contrasting AFM variations are a) to pursue models with a greater bias toward explanatory simplicity as there are many others, as indicated above, pursuing more complex models and better prediction with relatively less concern for interpretability and downstream application and b) to more directly build off the prior work on transfer that used AFM variations but left an open question as described above[4].

3. METHODS

The component theory of transfer uses a matrix to map knowledge components to items. The strong model suggests a single explanation for both difficulty and transfer but sacrifices some predictive power. The weak model theory offers better prediction but is less explanatory. We investigate an alternative model that combines benefits of the strong and weak models in a hierarchical fashion. We hypothesize that this alternative model will provide the explanatory power of the strong model without losing the predictive power sometimes better displayed by the weak model. We seek the interpretability and application benefits of explaining both difficulty and transfer using KCs, but without losing predictive power as has been observed with a strong component model or AFM.

3.1 Datasets

A variety of datasets representing four domains including geometry, algebra, English articles and statistics, with different task ordering approaches, were selected from LearnLab's DataShop [2], an open repository for diverse educational domain data. The datasets had different characteristics (e.g., number of students, number of KCs, etc,), which have been reported previously [4].In the educational technology applications that students used in producing this data, students solve problems or answer questions sometimes with multiple steps with feedback. Each evaluated step is considered a task or assessment "item" and can be labeled with one or more knowledge components (KC). Each dataset had multiple knowledge component models associated with it, where each model represents a different mapping from steps/items to skills/KCs. In this paper, for each dataset we used best KC model generated by LFA. The best was selected using the lowest root mean squared error (RMSE) on item-blocked cross-validation (explained below). We compared three different statistical models across six datasets.

3.2 Metrics of predictive accuracy

To evaluate predictive accuracy, we used five independent runs of 10-fold cross-validation (CV) using three variations of how the folds are produced, randomly, blocked by item, blocked by student, as per standard practice in LearnLab's DataShop[2]. We explain the item-blocked and student-blocked approaches next. The prior work[4] compared the component strong model and component weak model by creating folds in CV and blocking data records either by item or by student. In item-blocked CV, on each iteration, all data for an item is either in the training set or test set, but never both. In that prior work, the prediction fit of the weak models was consistently worse than the strong models when tested for generalization to new items, that is, via item blocked CV. This result can be explained by noting that in the weak model item difficulty estimates are not available for predicting test set data. The weak model relies only on overall difficulty, as well as student proficiency and KC learning rate, to predict on test set. In contrast, the strong model can use KC difficulty, as well as student proficiency and KC learning rate, to predict on test set. At the same time, it is important for models to generalize to new students and thus testing them via student-blocked CV is also sensible. In this case, prior work[4] demonstrated that the strong models were consistently worse than the weak when tested for generalization to new student, that is, via student-blocked CV. This observation leads to a central question of this paper: Can we address this prediction fit limitation of the strong component model (AFM) without losing the explanatory power of the KC difficulty estimation?

3.3 Statistical Models

To fit the statistical models, we used a generalized linear mixed-effects model (lme4 package in R)[6] to specify both random and fixed effects parameters. All three models set student proficiency as a random effect and learning rate as a fixed effect, thus leaving the difficulty parameter as the discriminant for prediction.

3.3.1 Strong component Model(AFM)

The strong component model, also known as the Additive Factors Model (AFM), is a logistic regression statistical model shown in

Table 1: A comparison of three cross validation results (random, student-blocked & item-blocked) using root mean square error across three component models (strong, weak & hierarchical) for six datasets. The RMSEs in bold indicate the best predictive models.The hierarchical model is the best predictor for all six datasets for random and student-blocked CV and for 2 of 6 datasets for item-blocked CV. Small differences are seen in the remaining 4 datasets in the item-blocked CV between the strong and hierarchical models.

	Component Strong (AFM)			**Component Weak**			**Hierarchical Model**		
Data	Random	Student-blocked	Item-blocked	Random	Student-blocked	Item-blocked	Random	Student-blocked	Item-blocked
Geom	0.3972	0.4068	**0.3990**	0.3970	0.4055	0.4158	**0.3955**	**0.4051**	0.4009
Stats	0.3253	0.3287	**0.3357**	0.3169	0.3212	0.3493	**0.3153**	**0.3198**	0.3379
Math	0.4380	0.4917	**0.4405**	0.4159	0.4716	0.4645	**0.4157**	**0.4714**	0.4413
Eng1	0.4078	0.4409	0.4157	0.4027	0.4388	0.4160	**0.4026**	**0.4374**	**0.4150**
Eng2	0.3747	0.4017	0.3843	0.3609	0.3898	0.3890	**0.3604**	**0.3894**	**0.3819**
Eng3	0.3892	0.4179	**0.3939**	0.3841	0.4127	0.3988	**0.3836**	**0.4121**	0.3942

R script in Equation 1. The response variable (correctness of student performance) is modeled as a function of the random effect for student proficiency combined with a fixed effect for knowledge component difficulty and a fixed effect for opportunity to practice each knowledge component. When KCs are modeled as fixed effects, the KC parameters estimates capture all the variance due to KC difficulty and there is no variance for items within the KC. The strong model uses parallel vectors of parameters with length equal to the number of KCs, thus explaining both difficulty and transfer using the same KCs.

$$correctness \sim (1|Student) + KC + KC : OppKC \quad (1)$$

3.3.2 Weak component Model(AFM')

The weak component model uses an item difficulty parameter to replace knowledge component difficulty found in the strong component model (see Equation 2). Unlike the strong model, the weak model provides a separate parameter for each item and, as such, does not provide a general explanation of difficulty, but merely a description of it. In this model, difficulty predictions (second term) are decoupled from transfer predictions (third term) as item is used for one and KC for the other.

$$correctness \sim (1|Student) + (1|Item) + KC : OppKC \quad (2)$$

3.3.3 Hierarchical Model(AFM'h)

The hierarchical model (AFM'h, Equation 3) models item difficulty through a hierarchical combination of KC-level and item-level estimates. Each item-level estimate is "pooled" within the KC it belongs and is thus constrained by the corresponding KC estimate. Item estimates are variations on the KC estimate and the model fit is penalized for item estimates away from zero (even as those estimates may improve correctness prediction). In machine learning terms, this constraint on item estimates is a kind of regularization. AFM'h provides an explanation of task difficulty in terms of knowledge components (as in AFM) but also provides an estimate of item difficulty (as in AFM').

$$correctness \sim (1|Student) + (1|Item/KC) + KC : OppKC \quad (3)$$

4. RESULTS

Figure 1a shows the root mean square error (RMSE) results from item-blocked CV across six datasets. The figure shows the weak model (see gray bars) is disadvantaged when generalizing across items as demonstrated by the height of the gray bars in comparison to the strong models (white bars) and hierarchical models (black bars). In all six datasets the weak models fared worse with item-blocked CV. Likewise, Figure 1b shows the RMSE for student-blocked CV where the component strong model (see white bars) fares poorly for all datasets in generalizing across students. These results confirm previous findings in which the strong model did better than the weak model in item-blocked CV (8 of 8 datasets) suggesting weak model has predictive disadvantages as well as explanatory ones[4]. Conversely, [4] found the weak model did better than the strong model in 7 of 8 datasets for student-blocked CV. Thus, there are predictive disadvantages of the strong model.

The important new result is that, while maintaining explanatory coherence and simplicity, the hierarchical model does not have either of these prediction disadvantages. See Table 1 for details. For both item-blocked and student-blocked CV, the prediction fits of the hierarchical model are never the worst. For student generalization (the student-blocked CV), 6 of 6 datasets have better prediction compared with the strong and the weak models and 2 of 6 for item generalization. They are essentially tied with the strong model in item generalization in the other cases. The differences slightly in favor of the strong model are in the Geom, Stats, Math, and Eng3 datasets. The pattern of results of random cross-validation, where data records are randomly assigned to CV folds irrespective of student or item tags, is highly similar to the pattern of results of student-blocked CV. Namely, the strong model is consistently the worst in prediction fit and the other two models are essentially tied.

5. DISCUSSION

This project advanced from the work done by [4] who provided clear evidence against the faculty theory, but identified an open issue about how best to implement the component theory. We hypothesized that combining the predictive power of a weak component model with the explanatory power of a strong component model would capture the best features of both models. Indeed from our results, the hierarchical models have as good or better prediction performance compared with the other two. The hierarchical model removes the prediction disadvantages of the strong model (for both student and item generalization) and both the explanatory disadvantage of the weak model and its prediction disadvantage on item generalization. Hence, we find that the hierarchical model gets the best of the both worlds: good prediction and good explanation.

One limitation of this study is that all the KC models used were single-KC models where each item is labeled by just one KC. Future work could attempt extend to models with multi-KC labeled items. Another future work possibility is to explore the use of the item random effect estimates of the hierarchical model as a better guide for KC model search than in recommended practice[9]. By comparing features of the hard items with those of the easy items, an analyst can hypothesize possible hidden skills and test whether associated KC relabeling produces better prediction fit. Current recommended practice relies on item means that may be biased estimates of item difficulty. These estimates are prone to inaccuracy particularly when there are a limited number of data points for an item. This situation occurs frequently in early system design and testing, which is just the point in system development when

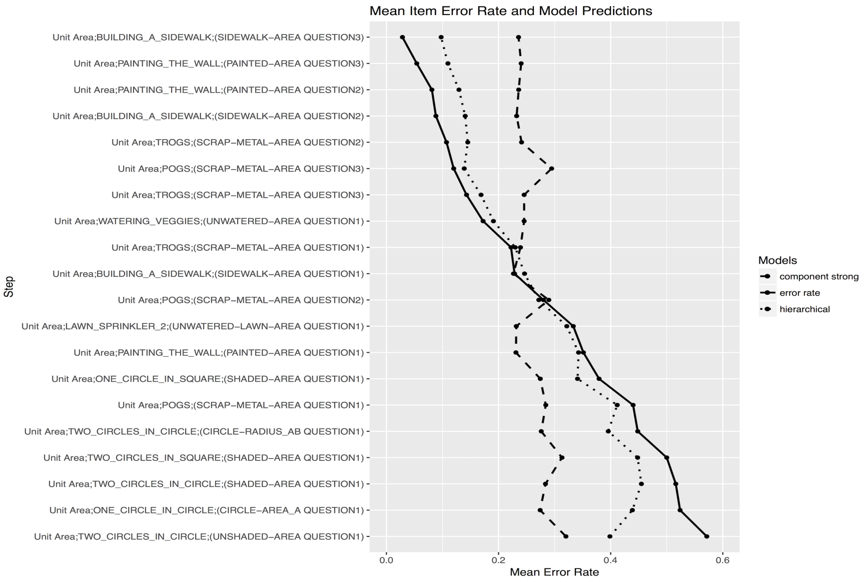


Figure 2: The hierarchical model (dotted line) better predicts item means (solid line) than the strong model (dashed line), but not perfectly so. Differences may indicate cases where the item mean is miss-estimating actual item difficulty perhaps because of limited data. An analyst trying to improve a KC model may be better guided by the hierarchical model predictions than by the simple means.

KC model testing and improvement is most needed. Low item frequency also occurs in systems that automatically generate a wide variety of items (e.g., with random numbers in math). The hierarchical model provides a less biased estimate of item difficulty, which is more robust to small samples given that it is influenced by data on other items within the same KC. The KC estimate serves as a Bayesian prior for all items embedded within it.

Figure 2 shows a performance profiler displaying 20 steps/items labeled by a suspect KC in an early non-optimized KC model for the Geom dataset. The items are sorted by mean error rate as displayed in the solid line. In dashed line are the error predictions of the strong component model (AFM) which are particularly poor because an early and non-optimized KC model is being used. The hierarchical model predictions (in dotted) are closer to the mean error rate, but are importantly different. In particular, the last row in Figure 2 is the item with the highest mean error rate, but the hierarchical model suggests that it may not be so hard. Since analysts trying to improve a KC model are looking to identify possible knowledge demands that differentiate harder and easier items, it may be helpful for them to not be deceived by possible miss-estimates of difficulty resulting from using item means.

6. CONCLUSIONS

The search for highly predictive statistical models is a major focus of educational data mining and data mining more generally. This search is often pursued with less attention to the explanatory power of the models. Good explanatory models provide both scientific insight about the nature of learning and interpretable implications for improvement in educational interventions. This paper provides a model case, which we hope others will follow, of seeking a method that provides both predictive accuracy and explanatory power.

7. ACKNOWLEDGMENTS

This work was supported in part by a National Science Foundation grant (ACI-1443068). We thank Hui Cheng for help with programming and data analytics support.

8. REFERENCES

[1] S. M. Barnett and S. J. Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612, 2002.

[2] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, 43, 2010.

[3] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[4] K. R. Koedinger, M. V. Yudelson, and P. I. Pavlik. Testing theories of transfer using error rate learning curves. *Topics in cognitive science*, 8(3):589–609, 2016.

[5] R. Liu and K. R. Koedinger. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, 9(1):25–41, 2017.

[6] P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[7] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[8] M. K. Singley and J. R. Anderson. *The transfer of cognitive skill*. Number 9. Harvard University Press, 1989.

[9] J. C. Stamper and K. R. Koedinger. Human-machine student model discovery and improvement using datashop. In *International Conference on Artificial Intelligence in Education*, pages 353–360. Springer, 2011.

[10] M. Wilson and P. De Boeck. Descriptive and explanatory item response models. In *Explanatory item response models*, pages 43–74. Springer, 2004.