TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 8 (2016) 589–609

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12208

This article is part of the topic "Discovering Psychological Principles by Mining Naturally Occurring Data Sets," Robert L. Goldstone and Gary Lupyan (Topics Editors). For a full listing of topic papers, see: http://onlinelibrary.wiley.com/doi/10.1111/tops.2016.8.issue-3/issuetoc.

Testing Theories of Transfer Using Error Rate Learning Curves

Kenneth R. Koedinger,[a] Michael V. Yudelson,[b] Philip I. Pavlik Jr.[c]

[a]*School of Computer Science, Carnegie Mellon University*
[b]*Research Department, Carnegie Learning, Inc.*
[c]*Institute for Intelligent Systems and Psychology, University of Memphis*

Received 31 October 2013; received in revised form 3 July 2014; accepted 3 July 2014

Abstract

We analyze naturally occurring datasets from student use of educational technologies to explore a long-standing question of the scope of transfer of learning. We contrast a faculty theory of broad transfer with a component theory of more constrained transfer. To test these theories, we develop statistical models of them. These models use latent variables to represent mental functions that are changed while learning to cause a reduction in error rates for new tasks. Strong versions of these models provide a common explanation for the variance in task difficulty and transfer. Weak versions decouple difficulty and transfer explanations by describing task difficulty with parameters for each unique task. We evaluate these models in terms of both their prediction accuracy on held-out data and their power in explaining task difficulty and learning transfer. In comparisons across eight datasets, we find that the component models provide both better predictions and better explanations than the faculty models. Weak model variations tend to improve generalization across students, but hurt generalization across items and make a sacrifice to explanatory power. More generally, the approach could be used to identify malleable components of cognitive functions, such as spatial reasoning or executive functions.

Keywords: Naturally occurring data; Faculty theory of transfer; Component theory of transfer; Model comparison; Learning curves

Correspondence should be sent to Kenneth R. Koedinger, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Office 3601, Pittsburgh, PA 15213. E-mail: koedinger@cmu.edu

1. Introduction

The question of how far learning in one context transfers to performance in another is quite old, but one of fundamental importance both to understanding the nature of human intelligence and to educational application (cf., Barnett & Ceci, 2002; Chen & Klahr, 2008; Day & Goldstone, 2012; Nokes-Malach & Mestre, 2013). In this paper, we use multiple datasets to compare alternative answers to this question of what is the scope or grain size of transfer. One simple take on this issue views the mind as a muscle and suggests that as one gets more experience, one gets smarter. Learning Latin, geometry, chess, or programming increases one's intelligence. A somewhat more refined version suggests multiple mental muscles or faculties of mind (cf., espoused by Thomas Reid in the 18th century as described in Yaffe & Nichols, 2009) and that transfer occurs by strengthening the relevant faculty. The associated doctrine of formal discipline (cf., Singley & Anderson, 1989) suggests, for instance, that a course in Euclidean Geometry exercises the reasoning faculty and thus would well prepare a law student because that faculty is employed in legal reasoning.

A starkly contrasting view to this faculty theory of transfer is Thorndike's identical elements theory of transfer (Thorndike, 1906). It suggests that transfer will only occur across tasks that require the same stimulus-response bond. The scope or generality of the "stimulus" is not well specified in the Thorndike's theory, but his empirical demonstrations suggest that the scope is quite narrow (Thorndike, 1922). For example, while students' error rate is just 6% on "What is the square of $x + y$?" (T1), it rises to 28% on the apparently isomorphic task "What is the square of $b_1 + b_2$?" (T2). The scope of a stimulus is presumably a bit broader than a specific task such that the same stimulus-response bond would be relevant to both T1 and a close isomorph such as "What is the square of $y + z$?" but is still narrow enough to reduce transfer to T2.

An intermediate view is what we call the "component theory of transfer," which suggests that transfer is achieved through cognitive procedures and structures that are more general than stimulus-response bonds, but more specific than domain faculties. A version of this component theory of transfer is expressed in Singley and Anderson (1989) where the declarative and procedural knowledge representations of the ACT-R theory are proposed as the components of transfer. Singley and Anderson do not so much resolve the question of what is the grain size or scope of transfer, but they suggest a representational tool for cognitive scientists to precisely specify components that predict the scope of transfer. They focus primarily on what they call as procedural transfer[1] (more robust, longer-term transfer) and on the production rule as the component of transfer: "transfer between tasks should depend on the degree to which they share common productions." In technical terms, the exact scope of transfer is specified by the set of tasks to which the variables in the if-part of a production rule can be bound. The specifics of the production rule notation are *not critical* to the more general notion of a component theory of transfer. Any knowledge representation (e.g., schemas or relational networks) that specifies the general conditions under which mental or physical responses are made is sufficient. What

is critical to the component theory of transfer is the notion of a unit of transfer. The breadth of application of the units (or "components" in the language we use in the paper) can vary. Here, faculty theory can be considered an extreme (or degenerate) instance of the component theory where there is a single component that has a complete general breadth. However, the key claim of the component theory is that the most or more accurate (i.e., better predicting) transfer model is going to have many components (not just one) per domain. These components will have a significantly more narrow breadth than the single component in the faculty theory. The single component in the faculty theory is relevant to all tasks in a domain, whereas each of the many components in a component theory is relevant to a subset of tasks (or task steps) in that domain. To qualify as a component theory (rather than as a faculty theory), there must be more than one component, and usually at least an order of magnitude more (e.g., at least 10).

A unit of knowledge transfer is often referred to as "knowledge component." A knowledge component (KC) is defined in the KLI Framework (Koedinger, Corbett, & Perfetti, 2012) as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks." Importantly, this definition grounds KCs empirically in terms of human task performance (cf., Anderson & Lebiere, 1998).

In addition to suggesting production rules, a tool for analyzing and predicting transfer, Singley and Anderson provide a useful review of empirical methods for testing claims about the scope of transfer. These methods tend to involve comparison of performance on two tasks or two sets of tasks, say A and B, where experimental designs are used to vary the ordering of A and B, for instance, comparing performance on B when, training on A occurs before it (AB), no training on B occurs before it (B or BA), or training on B occurs before it (BB). These all involve experimental manipulation and a pairing of tasks. Singley and Anderson also explore the use of learning curves to assess the scope of transfer, where task performance is naturally observed, not experimentally manipulated, and where performance on a large number and variety of tasks is observed. In this paper, we extend this learning curve analysis approach by including a broader set of comparisons and suggesting a more sophisticated statistical modeling approach.

Along with the empirical evaluation of (potentially atheoretical) proposals for the grain size of transfer, it would be ideal to have a theoretical mechanism that can predict the grain size of transfer or, equivalently, that can derive the scope of the conditions of KCs, from first principles. While some progress is being made on that front (e.g., Li, Stampfer, Cohen, & Koedinger, 2013), this paper focuses on an empirical mechanism that not only tests claims about the scope of transfer, but does so with now-prevalent and easy-to-access "found" data. This data come from natural student use of educational technologies as part of ongoing courses. Analysis of such data serves as a complement to the more costly (but worthwhile) experimental methods summarized by Singley and Anderson. The method we discuss is part of an approach to using found data to generate alternative KC models and find among them an empirically justified model for the scope of transfer (Koedinger, McLaughlin, & Stamper, 2012).

KCs serve two roles with respect to predicting human performance data. One role is in accounting for differences in task difficulty, and another role is accounting for transfer of

learning from one task to another. A *strong theory* does both at once—the KCs that explain task performance are also those used to explain transfer. A *weak theory* of transfer ignores the goal of explaining task difficulty and only uses KCs to explain transfer.

Stark contrasts are excellent to engage in debate, but our formulation allows the vast difference between the faculty theory and the identical elements theory of transfer to be considered along with a full range of component grain sizes. The faculty theory suggests a coarse grain component with many tasks (even domains) lumped together, whereas the identical elements theory suggests many finer grain components that split large categories of tasks (topics) into many groups of fewer tasks. At one extreme there is a single faculty, "intelligence," and transfer occurs across *all tasks.* At the other extreme, every task requires a different element and transfer only occurs at repeated exposures to the same task. As is so often the case, the truth likely resides somewhere between the extremes.

To address our question of the grain size of transfer of learning, we employ eight *naturally occurring datasets* (NODS) from student use of educational data technologies. These datasets are stored in DataShop, currently the world's largest open and free repository of educational technology data (Koedinger, Stamper, Leber, & Skogsholm, 2013). As with NODS in general, these datasets were not collected with our research question in mind. In particular, in most of the datasets, the order in which students experience tasks is not random, as might be ideal to explore our question of what is the grain size of transfer.

In what follows, we first describe an empirical strategy, based on error rate learning curves, for evaluating alternative theories of transfer. We next describe our methods for employing this strategy across the eight NODS. We then describe results of comparing alternative faculty and component models in their predictive accuracy and ability to explain transfer of learning. Finally, we discuss these results and their implications for advancing theory and models of the transfer of learning.

2. An empirical strategy for evaluating alternative theories of transfer

To test whether learning transfer is better accounted for by the faculty theory or the component theory, we need a strategy both for generating predictions from each theory and for testing those predictions against the data. We propose an approach that involves testing these theories across a selection of datasets in a variety of different domains. These datasets provide student performance data on tasks or items. The performance measure is error rate, that is, whether the student performs the item correctly or not (e.g., finds the correct area of a figure, picks the correct English article, clicks close enough to the correct location of a given fraction on a number line). All datasets come from student interactions with educational technologies whereby student attempts at items (including steps in complex problems) are followed by feedback and occasional instruction. Thus, student performance improves over time. In other words, we can use such data to generate learning curves in which error rates tend to decline as participants get more opportunities to practice.

Fig. 1 shows examples of error rate learning curves from one of these datasets. In general, learning curves represent a change in performance over learning time by relating some

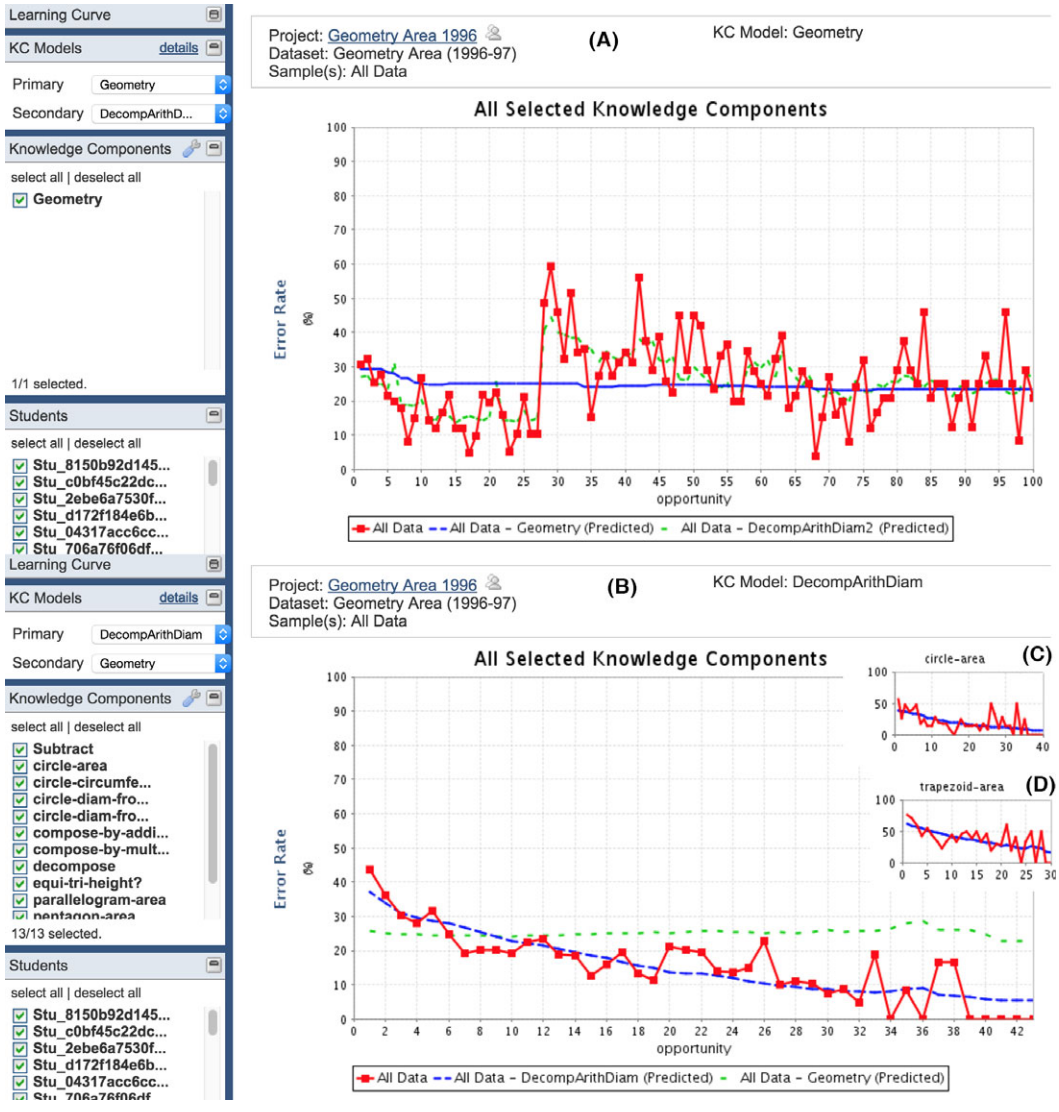

Fig. 1. Student learning curves (red solid lines) displaying the average error rate on geometry area items (vertical axis) over opportunities to practice (horizontal axis). In (A) the data are organized according to a *faculty theory of transfer*, whereby there is a "Geometry" faculty, and practice on any item should yield improvement on subsequent item performance. Transfer of learning is not apparent as the error rate does not regularly decrease as the opportunities to learn and practice increase. In (B) the same data are organized according to a *component theory of transfer*, whereby the same items are grouped within 13 *hypothesized* knowledge components (e.g., "circle-area," "circle-circumference," . . .) that students must learn to succeed on associated items. Transfer of learning is apparent in the decline in error rate with the increase in opportunities. Although (B) shows the average error rate across all components (and students), (C) shows the error rate for just the circle-area component and (D) for the trapezoid-area component (both averaged across students).

measure of performance (e.g., time to a correct answer or error rate) to some measure of learning time (e.g., number of trials or opportunities to practice). Fig. 1B, C, D illustrates typical learning curves (red lines) in that the error rate (*y*-axis) is decreasing as the opportunities to practice (*x*-axis) is increasing. The error rate is computed at each opportunity to practice by averaging student performance (right or wrong on an item labeled by a knowledge component) across all students and all knowledge components. For example, in Fig. 1B, the error rate on the first opportunity is about 44% indicating that of the 473 observations of students experiencing a knowledge component for the first time, in 206 cases the students failed to enter a correct response without help from the tutor (206/473 = 44%).

The solid red line of square points is the data (the blue and green dotted lines are statistical model predictions, which will be described later). In Fig. 1A, the opportunities to practice are counted according to a faculty theory of transfer, whereby every item a student performs is considered as an opportunity to practice or "exercise" the general faculty. Fig. 1A, however, does not look like a learning curve—the error rate is not going down in general as students are getting greater practice in the domain.

In Fig. 1B, the opportunities to practice are counted according to a component theory of transfer. For each item (often intermediate steps in a more complex problem), a knowledge component or skill label is assigned that is the hypothesized cognitive structure or process that must be acquired to perform correctly on this item. In this case, there are 13 component labels (shown in the Knowledge Component section on the left). With this categorization of items, we now see a learning curve (unlike Fig. 1A) whereby the error rate is decreasing as students have more opportunities to practice a particular component. Although Fig. 1B averages error rates across components (and students) at each opportunity, Fig. 1C, D shows learning curves for two of the thirteen components, circle-area (Fig. 1C) and trapezoid-area (Fig. 1D). These also indicate a decline in error rate as student opportunities to practice that component increase (note that the data are noisier at high opportunity counts because fewer students, and fewer components in the case of Fig. 1B, have data at these higher numbers of opportunities).

Consider the contrast between the erratic non-declining curve in Fig. 1A, when opportunities to practice are counted in terms of a *faculty* theory of transfer, and the smoother declining curve in Fig. 1B, when opportunities to practice are counted in terms of a *component* theory of transfer. This contrast is a springboard for using such data (from NODS) as a basis for an empirical test of the grain size of transfer. The visual comparison suggests that the component theory provides a better characterization of learning. We can make this intuition more precise through statistical modeling described below.

One may object that judging the faculty theory by the learning curve visualization in Fig. 1A is unfair. Perhaps there is faculty-level learning going on that is hiding within the variations in difficulty in the items. If one were to adjust for such item difficulty variation, this argument goes, perhaps transfer of learning would be apparent. This argument is the basis for considering the "weak" variation in the faculty theory introduced in statistical form below.

To apply the theories to each domain, we systematically develop statistical models in each domain capable of making predictions about student error rates across tasks and

opportunities to learn. We employ the notion of latent variables in a statistical model, a logistic regression model in particular, for this purpose (cf., Draney, Wilson, & Pirolli, 1996; Spada & McGaw, 1985). Table 2 shows four such models, two corresponding with the faculty theory and two with the component theory. In both cases, we explore both a strong version and a weak version following the useful distinction between descriptive and explanatory characteristics of statistical models introduced by Wilson and De Boeck (2004). In the strong versions, the theory is used to *explain both* task difficulty and transfer of learning across tasks. The weak versions maintain the goal of explaining transfer, but relax the goal of explaining task difficulty variations and merely *describe* it.

3. Methods and data

We use multiple datasets to explore the descriptive and explanatory quality of the space of variations of statistical models of transfer illustrated in Table 1.

All models in Table 1 capture variation in initial student proficiency (θ_i). The strong version of the faculty theory (AFM1) uses a single pair of parameters representing the faculty (indicated with the subscript 1) to explain both item difficulty (β_1) and learning transfer (γ_1) as a student gets successive opportunities to practice the faculty (T_{i1}). The weak version of the faculty theory (IRT + γ) does not attempt a general explanation of item difficulty (β_1 in the strong version), but instead has a separate parameter describing the difficulty of each item (β_j). The component theories of transfer use a matrix of component to item mappings (Q_{jk}) to indicate which knowledge component is required for each task item.

The strong version of the component theory of transfer (AFM) uses parallel vectors of parameters (length equal to the number of components) to explain both item difficulty (β_k) and learning transfer (γ_k) as students get successive opportunities to practice each component (T_{ik}). As with the weak faculty theory, the weak version of the component

Table 1
Statistical models of transfer predicting success[a] over time[b] based on student, item, and knowledge component factors[c]

	Faculty Theory	Component Theory[d]
Strong task difficulty (β) & transfer (γ) are *coupled*	$\theta_i + \beta_1 + \gamma_1 T_{i1}$ AFM1	$\theta_i + \beta_k Q_{jk} + \gamma_k Q_{jk} T_{ik}$ AFM
Weak task difficulty (β) & transfer (γ) are *decoupled*	$\theta_i + \beta_j + \gamma_1 T_{i1}$ IRT + γ	$\theta_i + \beta_j + \gamma_k Q_{jk} T_{ik}$ AFM'

Notes. [a]All models predict log-odds of success $ln(P_{ijT}/(1 - P_{ijT}))$ where P_{ijT} is the probability that student i is correct without assistance (1 or 0) on item j at practice opportunity T_{ik}.
[b]T_{ik} is the Tth opportunity of student i to perform (or further learn) component k.
[c]Parameter estimates for student proficiencies θ_i and item difficulties β_j are fit as random factors, while component difficulties β_k and learning rates γ_k are fit as fixed factors.
[d]Q_{jk} is a given matrix indicating which component(s) k are required to correctly solve item j.

theory of transfer (AFM') does not attempt a general explanation of item difficulty (β_k in the strong component theory), but instead has a separate parameter for each item (β_j). We also use a simple item response theory (IRT) model (the one-parameter Rasch model, see Wilson & De Boeck, 2004), which assumes there is no learning, as a baseline for contrast, particularly with IRT + γ.

The component theory of transfer, in fact, represents a space of possibilities depending on the components specified and their mapping to tasks. In terms of the statistical models, different Q-matrices are possible. In other work, we have explored this space using procedures for proposing Q-matrices that are executed by "hand" (Stamper & Koedinger, 2011) or machine (Koedinger, Corbett, et al., 2012; Koedinger, McLaughlin, et al., 2012). In this paper, we pick the best of each of these (details below) so that we have a hand-derived and machine-derived variation in each of AFM and AFM'.

3.1. Datasets

From over 700 datasets in LearnLab's DataShop,[2] we selected eight that were previously used to test an algorithm for semi-automated improvement of cognitive models called learning factors analysis (LFA; Koedinger, McLaughlin, et al., 2012). LFA combines a statistical model and combinatorial search over human-identified features. For statistical purposes, LFA represents a cognitive model of transfer as a *Q-matrix* (Tatsuoka, 1983)—a map of skills to tasks. It also uses a similarly structured *P-matrix* that specifies candidate features or "factors" for making cognitive model component distinctions including, for example, additional problem descriptors (e.g., "word problem" vs. "number problem") or contextualization for the use of the skills (e.g., "addition, small numbers" vs. "addition, large numbers"). In current practice, the P-matrix is created as a union of task factors introduced in all prior hand-constructed models as the basis for a logical generation of new KC models for statistical evaluation. Computationally, LFA is a search algorithm that generates candidate Q-matrices (KC models) by applying simple logical operations to sampled factors from the P-matrix. It evaluates these Q-matrices using AFM to find best fitting parameter estimates and uses Akaike Information Criterion or Bayesian Information Criterion (Burnham & Anderson, 1998) to guide the search and rank order the quality of the KC models it generates. As with other machine learning algorithms, the success of the LFA procedure depends on the quality of the input feature representation.

We should note that LFA has no strong bias or selection criterion that directly favors models that predict an "increase in performance." If there is learning in the data (e.g., which is typically verified with separate pre-post improvement data), then KC models that predict learning are likely to be favored by LFA—because they better fit the data—but the driver is data prediction, not any bias that favors bigger learning slopes over small ones.

The descriptive statistics for the datasets are given in Table 2 below. Table A1 in the Appendix provides further details on these datasets including the DataShop dataset number and names in DataShop of the chosen hand- and machine-made KC models. Across these datasets, there were different task ordering approaches used. Datasets 3, 4, and 7, all addressing English article learning, used multiple fixed orders based on an

Table 2
Summary of the datasets (in chronological order in which they were produced)

Data No.	Dataset	Domain	No. Students	No. Items	No. Data Points	No. KCs Hand	No. KCs Machine
1	Geometry Area (1996–97)	Geometry	59	139	5,388	12	18
2	Statistical Reasoning and Practice (Fall 2009)	College statistics	52	113	4,599	25	9
3	IWT Self-Explanation Study 1 (Spring 2009)	English articles	120	85	6,697	23	26
4	IWT Self-Explanation Study 2 (Fall 2009)	English articles	99	84	7,345	19	15
5	Assistments Math 2008–2009 Symb-DFA	Middle school math	318	64	9,340	5	7
6	Cog Model Discovery Experiment Spring 2010	Geometry	123	2,035	42,839	46	43
7	IWT Self-Explanation Study 3 (Spring 2010)	English articles	97	180	5,916	18	13
8	Improving Number Sense	Fraction number line	51	166	4,319	26	22

experimental design (Wylie, Koedinger, & Mitamura, 2010). Datasets 1 and 6 were on Geometry (e.g., Stamper & Koedinger, 2011) and used mastery learning algorithm to order tasks (the initial problems in dataset 1 are in a fixed order, though). In dataset 2, students saw a fixed order of tasks on a web page but were free to determine the order themselves (Lovett, Meyer, & Thille, 2008). In dataset 5, multiple counterbalanced orders were used (fixed within each) based on an experimental design (Koedinger & McLaughlin, 2010). Dataset 8 had random task selection after five fixed warm-up tasks (cf., Lomas, Forlizzi, & Koedinger, 2013).

Each selected dataset had multiple skill models associated with it, both hand-constructed and discovered using the Learning Factors Analysis (LFA) algorithm. From these models, for each dataset we chose two: the best hand-constructed (AFM-hand and AFM'-hand) and the best LFA model (AFM-machine and AFM'-machine). The best was selected using the lowest item-stratified cross-validation root mean squared error (RMSE). One additional selection criterion was that the matched KC models must label the same subset of items.[3]

3.2. Parameter estimation procedure

To fit statistical models, we used a modified LIBLINEAR tool (Fan, Chang, Hsieh, Wang, & Lin, 2008). LIBLINEAR is a library that supports fitting of various models, of which we were interested in L2-regularized logistic regression. We created a modification of this regression method that was computationally equivalent to a mixed-effect L2-penalized regression.[4] One shortcoming of LIBLINEAR package is that it does not supply fit model parameters with respective standard errors (and p-values), but this feature is not critical to our analyses.

3.3. Prediction fit evaluation via cross-validation

We performed cross-validation to further verify the results and reduce the danger of circularity in over-fitting arbitrary particulars of the data. Cross-validation helps us to make sure there is no strong bias or selection criterion that directly favors models that predict an increase in performance. We used a 10-fold cross-validation procedure to estimate the goodness of fit of the seven models on each of the eight datasets. To compute the cross-validated root mean squared error (RMSE) of the prediction, we iteratively built models on 90% of the data and produced predictions for the remaining 10%. Predictions were accumulated across all 10 folds and used to compute one RMSE value for the dataset/model tuple. During cross-validation, the overall student opportunity counts for the AFM1 and IRT + γ models, and skill opportunity counts for AFM and AFM' models were computed on the full dataset and were not recalculated after the data was split into the folds.

To produce the folds, we used two stratification methods: item-stratified and student-stratified. We are most interested in predicting future performance when the student sees new problems. For this purpose, item-stratification is the most relevant. Because the assignment of data to the 10% folds is random, we ran the folding procedure and subsequently cross-validated 20 times and based on the 20 RMSE values we computed the means and standard errors for each model/dataset pair.

4. Results: Explaining learning and prediction fit

4.1. Explaining learning

We first compare how the different statistical models of transfer capture learning in each dataset. It is important to note that there is independent evidence that students in these datasets are indeed learning. In most cases, this evidence is that students perform significantly better on post-tests after using the educational technology than they perform

Table 3
Median learning rates (learning curve slope in log-odds) across datasets and statistical models

	Faculty Models		Component Models			
Data#	Strong AFM1	Weak IRT + γ	Strong hd AFMh	Strong mc AFMm	Weak hd AFM'h	Weak mc AFM'm
1	0	0.004	0.069	0.116	0.081	0.092
2	0	0.002	−0.004	0.096	0.063	0.015
3	0	−0.032	0.086	0.106	−0.037	0.013
4	0	0.005	0.071	0.081	−0.059	0.002
5	0	0.014	0.082	0.144	0.055	0.095
6	0	−0.001	0.148	0.132	0.152	0.139
7	0	0.008	0.089	0.129	0.041	−0.002
8	0	0.009	0.091	0.133	0.020	0.200

on matched pre-tests before using the technology. See Stamper and Koedinger (2011) for geometry, Lovett et al. (2008) for college statistics, Wylie et al. (2010) for English, and Koedinger and McLaughlin (2010) for middle school math. In some cases, the evidence is better in second-half than first-half performance within the technology (cf., Lomas et al., 2013, for fraction number line). Table 3 is a summary of best estimates of learning rate parameters (γ_1 and γ_k) for the different models. As shown in the first column, the strong faculty theory model does not capture any learning: Practice opportunity slopes in the AFM1 models are always zero.[5] The weak faculty theory, ITR + γ, captures learning with positive learning rate slopes in six of the eight datasets. However, the slopes are small and sometimes negative (in datasets 3 and 6). The strong component theory models (AFMh and AFMm) have consistent positive slopes with one exception, the hand-constructed model for dataset 2. These slope values are also much higher than the faculty models. The slopes of the weak component models (AFM'h and AFM'm) are not as consistently positive, with negative median slopes for two of the hand-constructed models and one of the machine-constructed models. Overall, the component models, particularly the strong ones, indicate a greater amount of transfer of learning per opportunity to practice than do the faculty models.

We suggest two reasons for the less consistently positive slopes in the weak component models. First, the fact that our naturally occurring datasets (NODS) do not have full randomization of order of items may lead to a confounding of learning and item difficulty estimation. In weak models, variance that may be due to learning could instead be captured by the item parameter estimates to the extent that items are somewhat consistently positioned (earlier or later) in the curriculum. For example, an item that regularly appears near the end of the unit may, in fact, benefit from transfer of learning from items experienced earlier. Second, because the number of skills is less than the number of items, skill difficulties (in AFM) are estimated using more data points on average than are available for item difficulty estimates in AFM' (and IRT + γ). For both reasons, the use of item difficulty estimates may make the weak models more susceptible to overfit. Skill difficulty estimates may be more reliable, and the strong models may be more robust to the order bias in our NODS.

4.2. Prediction fit by dataset

The results of the prediction fits of faculty models (AFM1 and IRT + γ) and component models (AFMh, AFMm, AFM'h, and AFM'm) using root mean squared error (RMSE) on the test sets in item-stratified and student-stratified cross-validation (see Tables A2 and A3 in Appendix) show that predicting models are exclusively component models for 14 of 16 comparisons (8 datasets $\times$ 2 metrics). For the other two comparisons (all student-stratified), component models are in the set of best predicting models along with the weak faculty theory (IRT + γ) as well as the simple IRT model, which predicts no transfer at all.

Only once is a component model is the worst set of the 16 dataset-metric comparisons (i.e., the rows in Tables A2 and A3)—see the dataset 5 row in the item-stratified

cross-validation. One of the two faculty models is in the worst set in all 15 (of 16) other comparisons. In 12 comparisons, the strong faculty model is among the worst and in seven the weak faculty is among the worst.

4.3. Prediction fit ranking across datasets

The average rank data in Table 4 provide another way to compare models. It lists average model ranks within each of the two stratifications (by item and student) as well as overall average rank across stratifications. Models are ordered by overall rank. We clearly see that component models are on top of faculty models overall. However, the weak faculty model (ITR + γ) does beat one of the four component models in the student-stratified ranking. Component models that have machine-generated skill models rank better than hand-generated models overall, and within weak and strong component model categories across stratifications. For item-stratified ranking, both strong component models have an edge over weak component models, and for student-stratified ranking, weak component models have an edge.

The lower prediction error of the weak faculty model (IRT + γ) in the student-stratified cross-validation is due to the advantage of having $_j$ parameters capturing item variability. In the case that the test fold of the data contains new students, a strong component model with hand-set skills (AFMh) has only the information about skills that it got from the training. Having item variability parameters gives the weak faculty model and the weak component model an edge in this situation. The strong component model with machine-discovered skills, however, beats the weak faculty model. Consistent with the goal of Learning Factors Analysis, the machine-discovered skills appear better suited to capture transfer across folds.

5. Discussion

The strong faculty theory can be clearly rejected, both because it does not capture any transfer of learning (the slopes are all 0) and because the prediction results are consistently outperformed by other models. The case for the weak faculty theory (IRT + γ) is better, but it is not great. It is never one of the best predicting models when generalizing

Table 4
Average ranks of the models across eight datasets within stratifications and overall

Model	Weak/Strong	Faculty/Component	Item-Stratified	Student-Stratified	Overall
AFMm	Strong	Component	1.125	3.625	2.375
AFM'm	Weak	Component	3.125	1.750	2.438
AFM'h	Weak	Component	4.125	2.000	3.063
AFMh	Strong	Component	2.125	4.250	3.188
IRT + γ	Weak	Faculty	5.500	3.375	4.438
AFM1	Strong	Faculty	5.000	6.000	5.500

across items (i.e., in item-stratified cross-validation), but it ties for best on two of the eight datasets (#2 and 5) when generalizing across students (i.e., in student-stratified cross-validation). The estimated learning slopes (γ) in these two cases are positive; however, the 1PL IRT model, which has no learning slope, has equal predictive power. In other words, there is no statistically reliable evidence that the learning transfer apparently captured in these weak faculty theory models is real.

Why does the weak faculty model (and 1PL IRT itself) sometimes yield predictions as good as the ones of the component models? While component models are consistently better than faculty models on item-stratified cross-validation, they do not have the same consistent advantage on student-stratified cross-validation. Because the theory of transfer is most fundamentally about how learning transfers across *items*, item-stratification is arguably more relevant to testing a theory of transfer. Nevertheless, the question is worth investigating. A hierarchical statistical model that includes a global faculty and allows for item variation within knowledge components, while being harder to interpret, may be a more accurate statistical modeling approach and is worth future exploration, but it is beyond the scope of this paper.

Overall we can conclude that when comparing the stronger versions of the faculty theory (AFM1 model) and component transfer theory (AFM model), the evidence is clearly in favor of the strong component theory over the strong faculty theory. For the weak versions of the theories, the evidence is less decisive but weighs clearly in favor of the weak component theory (AFM') over the weak faculty theory (IRT + γ).

5.1. Strong versus weak component models

The results of comparing the strong versus weak component models on prediction accuracy are mixed. The best models are exclusively strong component models in item generalization and nearly exclusively weak component models (7 out of 8) in student generalization. Weak models do more poorly than strong models in item generalization because, across folds, items in the test set do not appear in the training set and thus the relevant item parameters are left to take on a default value. In strong models, the KC parameters fit to items in the training set are used for (transfer to) the same-labeled items in the test set.

Why weak component models have higher predictive accuracy than strong component models in student generalization is not clear, but it may be because items have relatively consistent positions during training in many datasets. Alternatively, there may well be significant item variability within sets of items labeled by the same component. For example (from dataset #8), even though 1/10 and 1/8 may draw on the same core knowledge (they are both unit fractions near 0) and exhibit mutual transfer, accurately placing 1/10 on a number line appears to be consistently easier than 1/8. As mentioned above, future research should explore a hierarchical model where item difficulty is pooled within component difficulty.

Turning to explanatory adequacy, the strong component models appear to better capture learning in that their learning rate (slope) estimates are more consistently positive

(with only one case of 16 where the median slope is negative) and tend to be higher (11 of 16 cases) than the median slopes of weak component models. Most important, a strong component theory has greater explanatory power in that it provides an explanation of variation in task difficulty that is one in the same with its explanation of variation in transfer of learning. The weak component theory, which is much less parsimonious (because there are many more item parameters than KC parameters), does not provide an explanation of task difficulty.

That variation in task difficulty should co-vary with learning transfer is a non-trivial and scientifically important notion. It is one that is consistent with general cognitive theories of learning and performance where cognitive structures and processes are hypothesized to be acquired and to account for human performance in tasks (e.g., ACT-R, Soar, Icarus, SME, LISA). For example, SME and LISA are component models in that analogies produce localized changes to knowledge structures—new relations or role fillers are added. Although these cognitive structures and processes vary in their character from theory to theory (e.g., working memory chunks, production rules, schemas, relational hierarchies), these theories all suggest that different kinds of tasks may require different components. These component differences, then, can explain why some tasks are harder than others (e.g., because one task requires components that another does not). Further, the success or failure of transfer from learning one task to performance on another can be explained by whether the components needed for those tasks are completely, partially, or not in common. In other words, these theories that are consistent with the strong component theory of transfer claim that task difficulty predictions and transfer predictions should derive from the same source.

In contrast, neural networks have a faculty character to the extent they use distributed representations, whereby every training task/example produces global/holistic changes in the network, not local ones. It may be, however, that such models often have a component character whereby neural network experience with "sing → sang" transfers to better performance on "ring → rang" but not on "jump → jumped."

5.2. Some limitations of the naturally occurring datasets (NODS)

One complication in visualizing and fitting slope parameters is that in some of our NODS, the educational technology system implements an adaptive mastery algorithm, and better students tend to master skills with fewer opportunities. The two datasets (1 and 6) from the Geometry Cognitive Tutor have this feature. In such datasets, the number of students contributing to the error rate average decreases as the number of practice opportunities increases. With better students mastering a component skill earlier, the remaining students are weaker ones. Thus, not only do learning curves get noisier for the higher ranges of opportunity counts, but they also may increase in error rate. Murray et al. (2013) provide a method for addressing this mastery-based selection bias. They demonstrated (in Bayesian models of learning not AFM-style models) that if learning curves are not aligned by first opportunity but by predicted mastery (often the last opportunity), learning slope estimates increase, indicating, they argue, a reduction in this bias.

Unlike the Bayesian network model used to model learning in that work, the AFM model used here has student specific parameters (θ_i) that at least somewhat adjust for this mastery-based selection bias. Most importantly for the purposes of this paper, it does not appear that the results for datasets coming from mastery-based tutors (1 and 6) are different from the others.

When examining student-stratified cross-validated RMSE, we saw, in two of the eight cases, that the prediction fit of the component models (AFM'h and AFM'm) was matched by the weak faculty theory model (IRT + γ) and IRT. The models that have an item intercept may take advantage of having more parameters that span training and testing chunks in cross-validation. Variance that may be due to learning may instead be captured in higher success parameter estimates for items that tend to be experienced later in training. This potential order bias is a consequence of the fact that the task order in the NODS we used was not fully randomized (and even fixed in some datasets). Note that this kind of order bias is not an issue (or, at least, much less of an issue) for the strong models because observations associated with a component or faculty are distributed widely across the training sequence even in cases where there is no order variation across participants. In other words, the prediction fit of the weak models may be over-estimated in our sample of NODS.

5.3. Empirically identifying the components of transfer

Given the evidence in favor of a knowledge component theory of transfer, it is natural to turn our attention to how a component model is formulated and improved. How can a proposed component model be enhanced? One strategy for exploring such improvement follows the logic illustrated above in Fig. 1. Fig. 1A shows the inadequacy of a single component model via its failure to produce a smooth, declining learning curve. However, splitting that component into finer grained components, the 13 shown in Fig. 1B, does produce a smooth, decreasing learning curve. This same strategy can be employed to a hypothesized component that has a rough or non-declining learning curve (i.e., looks like Fig. 1A).

Consider Fig. 2, for example, which shows an expanded view of the learning curve of one of the 13 components (the one in Fig. 1C). While declining, it is rough, indicating unaccounted for variance in the data. That variance may reflect that the component analysis is wrong. More specifically, it may be incorrect to model all tasks involving direct use of the circle-area formula as requiring the same knowledge component. In fact, through the application of the learning factors analysis (LFA) algorithm discussed above, it was discovered that tasks in which the circle-area formula is applied in a "backwards" direction (i.e., given a circle's area, find its radius) are different, both in difficulty and in mutual transfer, from tasks in which the circle-area formula is applied in a "forwards" direction (i.e., given a circle's radius, find its area). The nuance of this discovery is more explicit in the context of the fact that this forward-backward distinction is not empirically justified (there is no difficulty difference or lack of transfer) for any of the other area formulas (e.g., parallelogram, trapezoid, pentagon). It can be explained with a particular knowledge component, namely, knowing when to employ the square root operation, which is the mental operation uniquely required for backward application of the circle-

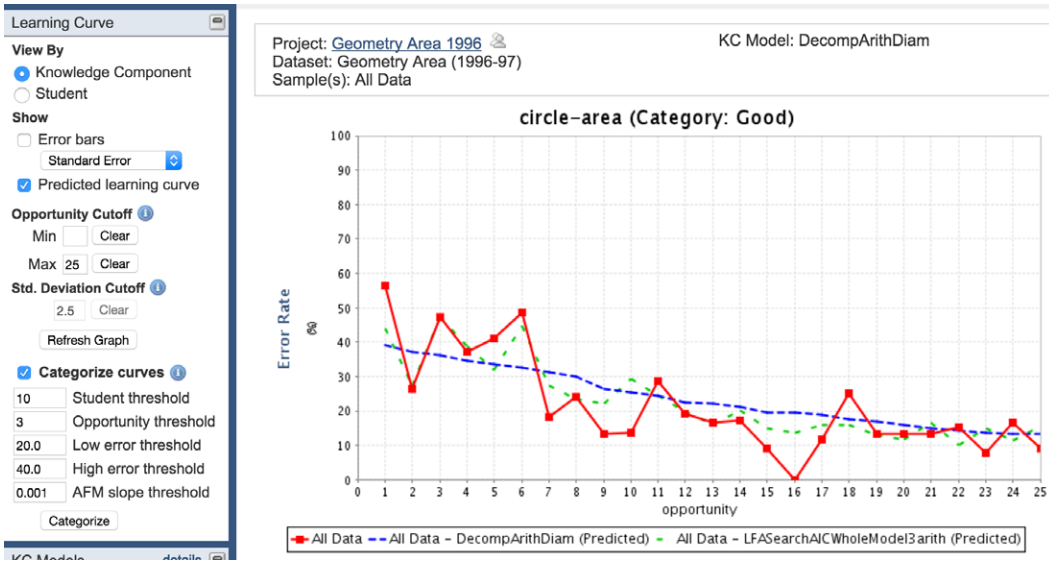

Fig. 2. An expanded view of the "circle-area" curve shown in Fig. 1C. Note the similarity in the comparison of AFM1 (Geometry) to AFM-hand (DecompArithDiam) in Fig. 1A. While going down, this curve (data in red solid line) still has some roughness indicating uncaptured variance. The machine-discovered model (LFA-SearchAIC, shown in green broader dashed line) better predicts the data because it makes a further distinction (separating whether r or A is given in applying the circle-area formula $A = \pi r^2$) that the hand model (DecompArithDiam; shown in blue tighter dashed line) does not.

area formula and none of the others. More generally, Koedinger, Corbett, et al. (2012) and Koedinger, McLaughlin, et al. (2012) demonstrate how the LFA algorithm aids in identifying the empirically justified grain size of transfer.

It is worth noting that while a component theory of transfer suggests knowledge components that are narrower in their breadth of applicability than a faculty theory suggests, it does not exclude the possibility of knowledge components that are broadly applicable. For example, learning to add two digit numbers can help in a variety of tasks: figuring out a tip at a restaurant, calculating a mean, and determining how many bacteria are on an agar plate. Students who have learned two-digit addition in one context are likely to transfer that knowledge to other contexts. This wide breadth, however, is not as broad as a general reasoning faculty or even an arithmetic faculty. Further, two-digit addition is not a single component—tasks of adding two digit numbers that involve a "carry" (e.g., 14 + 27) require a separate component from adding two digit numbers (e.g., 14 + 23) without a carry.

The strategy we are describing for statistical modeling and data-driven search for components of transfer is not limited to educational technology data. It could be applied to a multitude of questions about malleable components of intelligence and across what subsets of tasks transfer of learning occurs. For example, LFA could be employed to answer questions about what are the components of spatial reasoning, how malleable is it, and does training on particular spatial task domains (e.g., paper folding) transfer to all other spatial task domains (e.g., mental rotation, hidden figures), as per a faculty theory, or just

with those task domains or across certain subsets, as per a component theory (cf., Jee, Uttal, Gentner, Manduca, & Shipley, 2013). Similarly, LFA could also be applied to questions about what are the transferable components of executive functions (cf., Diamond & Lee, 2011; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). More broadly, a methodology for identifying the components of transfer of intellectual skills is of both scientific interest and practical importance.

6. Conclusion

Without replicating the simplicity of the faculty theory, a number of treatments of transfer suggest the possibility of transfer of a relatively broad variety (Barnett & Ceci, 2002; Bransford & Schwartz, 1999; Goldstone & Wilensky, 2008; Perkins & Salomon, 1994). In addition, there have been some substantial experimental demonstrations of broad transfer (Adey & Shayer, 1993; Chen & Klahr, 2008; Jaeggi et al., 2008; Klahr & Carver, 1988; Roll, Aleven, McLaren, & Koedinger, 2011; Schoenfeld, 1985; Schwartz & Martin, 2004). Some of these demonstrations were explicitly based on a detailed analysis of the components of desired thinking (e.g., a "cognitive model") and a corresponding development of instructional activities to aid the student mental construction of these components (e.g., Chen & Klahr, 2008; Klahr & Carver, 1988; Roll, Aleven, McLaren, et al., 2011; Schoenfeld, 1985). In other cases, follow-up analyses have explored component explanations of transfer results. For example, Koedinger and Wiese (2015) provide a component explanation for Adey and Shayer's (1993) evidence for long-term transfer from their reasoning-oriented instructional program in science to later math and reading assessments. In another example, Roll, Aleven, and Koedinger (2011) suggest alternative component explanations for how inventing activities lead to improved future learning (Schwartz & Martin, 2004).

In neither of these follow-up analyses is there much certainty that the component analyses proposed are the correct explanations. In both cases there are insufficient process level data (e.g., to produce learning curves) that might differentiate component explanations from other explanations. A significant challenge for the future is to find ways to combine innovative experiments toward achieving broad transfer with the collection (perhaps aided by educational technology) of the kind of fine-grained longitudinal process data needed to differentiate alternative explanations for how transfer is achieved.

Notes

1. They also discuss declarative transfer that is observed only in "the initial period" of training and "transfer among tasks [occurs] to the degree that the tasks share a common declarative base."
2. LearnLab's DataShop can be accessed at http://learnlab.org/datashop.
3. Not all recorded student actions are labeled with a KC as some may be considered trivial or irrelevant (e.g., entering a given value or clicking a done button). Differ-

ent KC models may make different assumptions about what actions are "items" and thus labeled.

4. We attempted to use the R statistical package "lme4" for fitting mixed-effect regression models, but some regressions were suffering from a loss of rank, which is not well-supported by lme4. Further, it ran too slowly for the largest of these datasets.
5. Strictly speaking, the parameter values are different from zero, but they are effectively zero up to fifth decimal point.

References

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. *Cognition and Instruction*, *11*(1), 1–29.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(6), 612–637.

Bransford, J. D., & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad, & P. D. Pearson (Eds.), *Review of research in education* (vol. 24, pp. 61–100). Washington, DC: American Educational Research Association.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach*. New York, NY: Springer-Verlag.

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (vol. 36, pp. 419–470). Amsterdam: Elsevier.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153–176.

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, *333*, 959–964.

Draney, K., Wilson, M., & Pirolli, P. (1996). Measuring learning in LISP: An application of the random coefficients multinomial logit model. In G. Engelhard, & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 195–218). Norwood, NJ: Ablex.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*(2008), 1871–1874.

Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *Journal of the Learning Sciences*, *17*(4), 465–516.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833.

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., & Shipley, T. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, *14*(2), 175–187.

Klahr, D., & Carver, S. M. (1988). Cognitive objectives in a LOGO debugging curriculum: Instruction, learning, and transfer. *Cognitive Psychology*, *20*(3), 362–404.

Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798.

Koedinger, K. R., & McLaughlin, E. A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 471–476). Austin, TX: Cognitive Science Society.

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated student model improvement. In Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. (Eds.), *Proceedings of the 5th international conference on educational data mining*. (pp. 17–24). Chania, Greece.

Koedinger, K. R., Stamper, J. C., Leber, B., & Skogsholm, A. (2013). LearnLab's datashop: A data repository and analytics tool set for cognitive science. *Topics in Cognitive Science*, *5*(3), 668–669.

Koedinger, K. R., & Wiese, E. S. (2015). Accounting for Socializing Intelligence with the Knowledge-Learning-Instruction Framework. In Resnick, L.B., Asterhan, C. and Clarke, S.N. (Eds.), *Socializing Intelligence through Academic Talk and Dialogue*. Washington, DC: American Educational Research Association.

Li, N., Stampfer, E., Cohen, W., & Koedinger, K. R. (2013). General and efficient cognitive model discovery using a simulated student. In M. Knauff, N. Sebanz, M. Pauen, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 894–899). Austin, TX: Cognitive Science Society.

Lomas, D., Forlizzi, J., & Koedinger, K. R. (2013). Optimizing challenge in an educational game using large-scale design experiments. In Mackay, W. E., Brewster, S., & Bødker, S. (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, ACM SIGCHI Conference, Paris, France, pp. 89–98.

Lovett, M., Meyer, O., & Thille, C. (2008). JIME - The open learning initiative: Measuring the effectiveness of the OLI Statistics course in accelerating student learning. *Journal of Interactive Media in Education*, *2008*(1), p.Art. 13. DOI: http://doi.org/10.5334/2008-14.

Murray, R. C., Ritter, S., Nixon, T., Schwiebert, R., Hausmann, R. G. M., Towle, B., Fancsali, S., & Vuong, A. (2013). Revealing the learning curves. In H.C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, AIED 2013, Memphis, TN, pp. 473–482.

Nokes-Malach, T. J., & Mestre, J. P. (2013). Toward a model of transfer as sense-making. *Educational Psychologist*, *48*(3), 184–207.

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In T. Hus.n, & T. N. Postlethwaith (Eds.), *International encyclopedia of education* (2nd ed., pp. 6452–6456). Oxford, UK: Pergamon Press.

Roll, I., Aleven, V., & Koedinger, K. R. (2011). Outcomes and mechanisms of transfer in invention activities. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2824–2829). Austin, TX: Cognitive Science Society.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Singley, K., & Anderson, J.R. (1989). The transfer of cognitive skill. Cambridge, MA: Harvard University Press.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169–194). Orlando, FL: Academic Press.

Stamper, J. C., & Koedinger, K. R. (2011). Human-machine student model discovery and improvement using data. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 353–360). Berlin: Springer.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Thorndike, E. L., (1906). *Principles of teaching.* New York: A.G. Seiler.

Thorndike, E. L., (1922). The effect of changed data upon reasoning. *Journal of Experimental Psychology*, *5*, 33–38.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer-Verlag.

Wylie, R., Koedinger, K. R., & Mitamura, T. (2010). Analogies, explanations, practice: Examining how task types affect second language grammar learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the international conference on intelligent tutoring systems* (pp. 214–223). Heidelberg: Springer.

Yaffe, G., & Nichols, R. (2009). Thomas Reid, *The Stanford Encyclopedia of Philosophy* (Winter 2009 Edition), E. N. Zalta (ed.). Available at http://plato.stanford.edu/archives/win2009/entries/reid/. Accessed April 7, 2016.

Appendix

Table A1
Dataset identifiers in the PSLC DataShop and knowledge component (KC) model names

Data No.	Dataset Name	Dataset ID	Hand-Constructed KC Model Name	Machine-Discovered KC Model Name
1	Geometry Area (1996–97)	76	xDecmpTrapCheat	LFASearchAICWholeModel3
2	Statistical Reasoning and Practice (Fall 2009)	308	SubLO1	LFASearchModel1
3	IWT Self-Explanation Study 1 (Spring 2009)	313	multiple KCs	LFASearchAICWholeModel1
4	IWT Self-Explanation Study 2 (Fall 2009)	372	Default_corrected	LFASearchAICWholeModel1
5	Assistments Math 2008-2009 Symb-DFA	388	KC7-split	LFASearchModel1
6	Cog Model Discovery Experiment Spring 2010	392	KTracedSkills	LFASearchAICWholeModel0
7	IWT Self-Explanation Study 3 (Spring 2010)	394	Default	LFASearchAICWholeModel1
8	Digital Games for Improving Number Sense	445	Item_add_Click_Type	LFASearchAICModel0

Table A2
Root mean squared error (RMSE) on test set prediction via cross-validation. Item-stratified cross-validation. The standard deviation of the mean RMSE reported is at most 0.007 and is 0.001 on average

	Faculty Theories		Component Theories				
Data#	Strong AFM1	Weak IRT + γ	Strong hd AFMh	Strong mc AFMm	Weak hd AFM'h	Weak mc AFM'm	Baseline IRT[b]
1	0.428	*0.431*[a]	0.404	**0.401**[a]	0.421	0.419	0.430
2	*0.370*	*0.371*	0.363	**0.352**	0.367	0.366	0.370
3	0.444	*0.448*	**0.428**	**0.428**	0.442	0.437	0.444
4	0.413	*0.420*	**0.392**	**0.392**	0.405	0.404	*0.420*
5	0.468	0.467	**0.454**	**0.453**	*0.472*	0.469	0.467
6	*0.366*	*0.367*	0.335	**0.333**	0.348	0.347	*0.367*
7	*0.423*	*0.424*	0.406	**0.398**	0.413	0.407	*0.424*
8	*0.456*	*0.456*	0.449	**0.438**	0.448	0.443	*0.456*

[a]The standard error of RMSE across 20 iterations is used to determine groups of winning models (bold) and of losing models (italic).
[b]IRT results provide a baseline for reference.

Table A3
Root mean squared error (RMSE) on test set prediction via cross-validation. Student-stratified cross-validation. The standard deviation of the mean RMSE reported is at most 0.002 and is 0.0005 on average

	Faculty Theories		Component Theories				
Data#	Strong AFM1	Weak IRT + γ	Strong hd AFMh	Strong mc AFMm	Weak hd AFM'h	Weak mc AFM'm	Baseline IRT[b]
1	*0.433*[a]	0.409	0.409	0.407	**0.406**[a]	**0.406**	0.411
2	*0.371*	**0.338**	0.343	0.347	**0.337**	**0.337**	**0.338**
3	*0.469*	0.444	0.448	0.441	0.442	**0.439**	0.444
4	*0.436*	0.397	0.409	0.411	**0.392**	0.395	0.397
5	*0.502*	**0.474**	0.492	0.492	**0.474**	**0.474**	**0.473**
6	*0.370*	0.339	0.344	0.342	**0.336**	0.337	0.340
7	*0.444*	0.415	0.427	0.422	**0.414**	**0.413**	0.415
8	*0.501*	0.490	0.490	**0.486**	0.491	0.490	0.490

[a]The standard error of RMSE across 20 iterations is used to determine groups of winning models (bold) and of losing models (italic).
[b]IRT results provide a baseline for reference.