All in the (word) family: Using learning decomposition to estimate transfer between skills in a Reading Tutor that listens

Xiaonan ZHANG, Jack MOSTOW, Joseph E. BECK
Project LISTEN, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. In this paper, we use the method of learning decomposition to study students' mental representations of English words. Specifically, we investigate whether practice on a word transfers to similar words. We focus on the case where similar words share the same root (e.g., "dog" and "dogs"). Our data comes from Project LISTEN's Reading Tutor during the 2003—2004 school year, and includes 6,213,289 words read by 650 students. We analyze the distribution of transfer effects across students, and identify factors that predict the amount of transfer. The results support some of our hypotheses about learning, e.g., the transfer effect from practice on similar words is greater for proficient readers than for poor readers. More significant than these empirical findings, however, is the novel analytic approach to measure transfer effects.

1. Introduction

One key problem in Intelligent Tutoring Systems is student modeling, i.e., inferring an individualized model that represents the student's state of knowledge. In this paper, we focus on one particular aspect of student modeling: estimating the transfer from learning one skill to similar skills. In other words, we investigate the extent to which practice on one skill improves another skill.

The amount of transfer reflects the student's mental representation of the skill, and may grow as the representation deepens. For example, when children first learn to read, they may well just treat each word as an independent entity; as they grow to be more advanced readers, they learn to represent a word as a sequence of phonemes, or as a base form with morphological changes [1]. Gradually in this way various words became representationally connected to each other. Transfer effects also affect the benefits of practice. Again take the example of learning to read. If there's no transfer between words, reading practice on one word is unrelated to any other word, and will contribute solely to learning that word; however, if the child has acquired the alphabetic principle, knowledge of one word should also transfer to words that are alphabetically similar. As a result, the child should make faster progress than he or she did when there was no transfer.

In this paper, we study transfer in the context of reading, and investigate whether the skill gained from exposure to a word transfers to similar words. The study is carried out in the context of Project LISTEN's Reading Tutor, an intelligent tutor that helps students learn to read [2]. It displays stories sentence by sentence on the computer screen, and listens to the student read aloud. The Reading Tutor uses Automatic Speech Recognition (ASR) to analyze the student's reading, and provides help when it detects a mistake or the student gets stuck. The student can also click for help when encountering difficulties. The time-aligned output from the recognizer is logged everyday into Project LISTEN's database. The Reading Tutor also logs its interactions with the student during the entire session, such as the sentence text that the student sees on the screen, or whether the student is helped on a specific word. As a result, the data available is copious and detailed, but noisy.

Our work is related to but different from some prior work. Both Chang [3] and Koedinger et al. [4] try to elucidate students' mental representation of the target skill, but by comparing competing models that differ in skill level and granularity. In contrast, we estimate *to what degree* a skill transfers to related skills, rather than focus on which single model fits best.

The rest of the paper is organized as follows. Section 2 describes our modeling approach in detail. Section 3 presents four analyses based on it, and their results. Section 4 discusses some limitations of our current work, and possible future directions. Section 5 concludes.

2. Approach

Our goal is to measure the amount of transfer from practicing one word to learning similar words. Sections 2.1-2.6 describe successive steps to achieve this goal.

2.1 Find a way to estimate the influence of various practice

We approach the problem with a relatively new method—learning decomposition [5]. It extends the classic exponential learning curve (Equation 1) by taking into account the heterogeneity of different learning opportunities for a single skill, as shown in Equation 2:

$$performance = A * e^{-b*t}$$

Equation 1. Exponential model

$$performance = A * e^{-b*(t1+\beta*t2)}$$

Equation 2. Learning decomposition model of practice

In both models, A measures students' performance on the first trial, and b represents the learning rate. The two models differ only in their counting of "trials." The simple learning curve uses a single variable t that pools all learning opportunities together. In contrast, learning decomposition partitions the t trials into $t1$ trials of type 1 and $t2$ trials of type 2, so as to differentiate one type of learning opportunity from the other by keeping count of each separately. The β parameter characterizes the relative effectiveness of type 2 trials compared to type 1 trials. For example, a β value of 2 would mean that practice of type 2 is twice as valuable as practice of type 1. The basic idea of learning decomposition is to find the weight β that renders the best fitting learning curve.

Equation 1 decomposes the learning opportunities into two types. More generally, we distinguish n types of trials by replacing t with $t_1 + \beta_2 t_2 + \ldots + \beta_n t_n$. That is, we just add a counter t_i for each type i, and a free parameter β_i to represent the impact of a type i trial relative to the "baseline" type 1, where $\beta_1 = 1$ by definition. We can number the types however we like, so we can freely choose which one to designate as the baseline by calling it type 1. Any non-linear regression package can fit the model to data; we use SPSS 11.0.

2.2 Define similarity

We hypothesize that exposures to a word may transfer to a similar word, but what does "similar" mean here? We operationalize "similar" as "sharing the same word root." By "word root," we mean the reduced form of a word after its morphological and inflectional endings are removed. We use Porter Stemmer [6] to extract word roots. Thus "dog", "dogs", and "dogged" are similar to each other because they all share the root "dog." In effect, we treat each word as a different skill, and focus on transfer between similar words.

It should be pointed out that there are many other ways to define words as "similar" (e.g., having at least 3 letters in common), and correspondingly many other ways to define what kind of transfer to model. However, the learning decomposition method is applicable no matter how similarity is defined, which is a great advantage of this approach.

2.3 Define the learning outcome

We need a performance variable that measures the quality of learning. As in the earlier work [5], we measure a student's performance on a word as the time to read it, capped at 3 seconds. The word reading time is computed from the Reading Tutor's log of the time-aligned ASR output. We also assign a reading time of 3 seconds to a word if the ASR rejects the word as misread or the student clicks on it for help.

Another concern is that our performance measure should distinguish short term retention from real learning. For example, suppose that a student sees two sentences with the word "dogs" in them, and then a sentence with "dog" shortly afterwards. We can imagine that the student should read the word "dog" fairly quickly, after having seen "dogs" twice. However, is it due to transfer from practice on similar words, or merely a temporary scaffolding effect? To avoid such ambiguities, we only use students' first encounters of a word root each day to estimate their knowledge, and exclude observations that may be contaminated by recency effects. However, subsequent encounters of the word root on the same day still count as practice opportunities. For clarity, we will use the term "outcomes" to refer to the subset of observations useful in estimating student knowledge, and "exposures" to refer to the entire set of observed practice opportunities. The model predicts only the outcomes, but all the exposures contribute to the model in calculating the amount (and types) of practice preceding each outcome. In the earlier example, the first encounter of "dogs" is an outcome for skill "dogs", while the two subsequent encounters count only as exposures for the skill "dogs" and "dog."

2.4 Partition the practice space

In our model, every word is a distinct skill. For each word, any previous encounter of the same word naturally constitutes a practice opportunity. Since we hypothesize that there is transfer from similar words, previous encounters of all similar words should also count as learning opportunities. Thus for every word, there is a unique practice space made up of two separate parts: exposures to similar words, and exposures to the word itself. However, not all practice opportunities are equal, even within these two parts. As reported in [5], massed practice (reading a word more than once in a day) is generally not as effective as distributed reading (reading a word on different days) for helping students to learn the word. With this in mind, we further distinguish between massed and distributed practice, and split all the practice opportunities for a word *w* into 4 non-overlapping types:

D (Distributed same-word exposure): see word *w* for the first time that day.
M (Massed same-word exposure): see *w* again on the same day.
DS (Distributed similar-word exposure): see its root for the first time that day, in some word similar to *w*.
MS (Massed similar-word exposure): see its root again that day, in some word similar to *w*.

Let's use an example to illustrate this partition. Suppose we are modeling the skill of reading "dog." In Table 1, the leftmost 3 columns list a student's successive encounters over two days of the word "dog" itself, and words similar to "dog" such as "dogs" and "dogged." The following four columns for each type (M, D, MS and DS) show the cumulative number of exposures of that type. The rightmost column shows whether the exposure also counts as an outcome. On Monday, the student sees the word "dog" for the first time, so the exposure is of type D. Then the student sees "dog" again, this time as a massed exposure for word "dog." On the 3rd exposure, the student encounters "dogs", which is similar to "dog." Since the student has already seen a word whose root is "dog" (namely "dog") that day, the counter for type MS is incremented. On the following Tuesday, the student sees "dogged"—another word similar to "dog", which counts as type DS since it's the first encounter that day. Another thing to note here is that all 5 exposures are practice opportunities, but only Exposure 1 and Exposure 4 count as outcomes, for the words "dog" and "dogged" respectively.

Table 1. Example of computing prior exposures to "dog"

Exposure No.	Day	Target Word	D	M	DS	MS	Count as Outcome?
1	Monday	dog	1	0	0	0	Yes
2	Monday	dog	1	1	0	0	No
3	Monday	dogs	1	1	0	1	No
4	Tuesday	dogged	1	1	1	1	Yes
5	Tuesday	dog	2	1	1	1	No

2.5 Decide how to aggregate across individuals

Now that we have described the basic components of the model, a question left unresolved is how to construct the model with data from multiple students. There are at least two straightforward solutions: 1. fit all the data to a single pooled model; 2. fit an individual model for each student, and then average the parameters in some way. We decide to go with the second option, for several reasons. First, individual estimates avoid the bias of aggregating data across students, since better readers typically generate more data, and are likely to differ a lot from poor readers in both initial knowledge and learning rate. Second, building student-specific models affords us flexibility in how we combine the models later in the analysis stage. For example, we may want to group students to make planned inter-group comparisons, as in [7], or cluster them after the fact.

However, partitioning exposures into different types already gives rise to a sparse data problem for rare types; estimates for individual students are even sparser, as reflected in the standard error output by SPSS for each parameter estimate. The obvious symptom of sparse data is an outlandish estimate or a large standard error. We deal with this issue by using the median (instead of mean) of the per-student estimates so as to deemphasize outliers. A less obvious symptom is a standard error of 0.0, which occurs (fortunately just occasionally) when the non-linear regression leaves a model parameter at its initial value due to lack of data. We deal with this issue by excluding these unchanged parameter estimates before computing the median.

2.6 Account for additional influences

Other factors also affect word reading time. First is word length: longer words generally take longer to read. Another factor is whether a student gets help on a word. We've adjusted reading time to 3 seconds for helped words, but we haven't taken into account the influence of previous help on that word or similar words. The number of times a student has gotten help on a word also reflects the difficulty of the word for the student. Incorporating all three factors leads to the full model shown in Equation 3.

$$performance = L * word\_length + A * e^{-b*(D+\beta_M*M+\beta_{DS}*DS+\beta_{MS}*MS+H*\#help\ on\ the\ word+HS*\#help\ on\ similar\ words)}$$

Equation 3. Full learning decomposition model for estimating transfer

In this model, we add the term word_length (number of letters in a word), and use L to weigh its relative impact. We also keep count of how many times the student requests help on the word as well as on similar words. Note that we only count student-initiated help, because the Reading Tutor could initiate help under a lot of (possibly inappropriate) situations so the influence of tutor help may be too complex to analyze. We set the weight of type D exposure to 1 so as to use it as a baseline. The three free parameters β_M, β_{MS}, and β_{DS}, reflect the impacts of exposures of type M, MS, and DS relative to the exposures of type D. The central task of this paper is to estimate and analyze the parameters β_M, β_{MS}, and β_{DS}.

3. Studies and Results

The data for this paper was collected by Project LISTEN's Reading Tutor during the school year 2003-2004, including 6,213,289 observations of 650 students' word encounters. To ensure that we measure learning rather than recency effects, the model predicts only outcomes as defined earlier. This filter reduces the data size to 2,711,618, though the other exposures still contribute to counting different types of practice. We further filter the data on several other conditions: 1. Exclude encounters of the 50 most frequent words. 2. Exclude all word encounters in stories the student has read before; 3. Consider only a student's first 20 exposures to a word (i.e., M+ D $\leq$ 20). 4. Exclude cases where a student's exposures to similar words (either massed or distributed) exceed 20 (i.e., we want MS $\leq$ 20 and DS $\leq$ 20). The rationale for these constraints is that under any one of the four conditions, the student should have gained too much familiarity with the word to exhibit any learning effect. 5. Exclude students who read fewer than 20 words in the Reading Tutor, because 20 data points is far from enough to give a reliable student-specific model. In fact, running non-linear regression with fewer than 20 data points for this model causes SPSS to abort the regression procedure. 6. Exclude students without paper test scores needed for some of our analyses. After filtering, we have 860,517 cases (trial outcomes), including 10,357 distinct words read by 346 students from Grade 1 to Grade 5.

We have some initial hypotheses before carrying out the experiment: 1. Students do benefit from exposures to similar words, perhaps not as much as they do from exposures to the word itself. 2. The transfer effect is stronger for students with higher initial reading skills. 3. Transfer is also greater for students with larger gains over the school year. We test our hypotheses in a series of analyses.

3.1 Are there transfer effects from practice on similar words?

We build a model in the form of Equation 3 for each student. Due to sparse data problems, the model construction fails for 18 students (one or more parameters aren't updated at all), so we take the medians only of the remaining 328 models as the overall parameter estimates. For each parameter, we also derive its two-sided 95% confidence interval using a non-parametric bootstrap test [8], by bootstrap sampling 10,000 times from the original estimates. The sample size is equal to that of the original set (328). Table 2 shows the result. The confidence intervals for different parameters vary in tightness, largely because some parameters are estimated from more data than others. The bottom row of the table shows the percentage of outcomes that contributes to each parameter estimate. For example, this percentage is 100% for L, because every word has a length, but only 7% for H, because students received help on only 7% of the words.

Table 2. Overall median parameter estimates (with 95% Confidence Interval)

Parameter	L	A	B	β_M	β_{MS}	β_{DS}	H	HS
Estimate	0.133 ±0.004	0.822 ±0.063	-0.082 ±0.014	0.171 ±0.069	0.063 ±0.114	0.551 ±0.126	-1.286 ±0.196	-0.665 ±0.257
% of outcomes	100	100	55.13	26.38	10.41	18.51	7.33	3.08

Overall, the model seems very reasonable. The positive value for L means that the reading time should grow with word length, on average 0.13 seconds per letter. The positive A represents students' initial reading time of a word, minus the approximated reading time for words of that length. The number seems a bit high at first look, but it's because we assign a value of 3 seconds to some words. Actually, in our dataset, the average reading time (after adjustments as described in section 2.3) for students' first word root encounters is 1.53 seconds. If we multiply estimated L by the average length of a word (5.5 letters in our data), and add the result (0.7315) to A, we get 1.5535, which is fairly close to the actual value of 1.53. The negative B shows the general trend that reading time decreases with more practice. Both H and HS are significantly less than 0. The negative coefficients for help presumably occur because help indicates that the student is having difficulty with the word, not that help is hurting the student.

Our result shows that M is significantly less than 1, which is consistent with the finding in [5] that massed practice is worth less than distributed practice. Practice of type MS and DS are also less effective than distributed reading of the same word, as expected. However, the impact of DS is quite notable ($\beta_{DS} > 0$ with $p<0.0001$). In fact, β_{DS} is even significantly larger than β_M ($p<0.0001$), which means that distributed practice of a similar word is more effective for learning a word than seeing the word itself again on the same day. In contrast, the impact of practice of type MS is negligible: the 95% confidence interval for MS straddles both sides of 0.

3.2 Higher reading proficiency, more transfer?

To investigate whether there is a positive relationship between initial reading skill and the amount of transfer, we divide all the students into 3 equal-sized bins: low proficiency, medium proficiency and high proficiency, according to their grade-equivalent pretest score on the Woodcock Reading Mastery Word Identification (WI) subtest [9]. Low proficiency students have a test score of less than 1.6. High proficiency students have scores greater than 2.5. Medium proficiency students are those with a test score in between. Then we calculate the median parameters for each subgroup of students. The result is shown in Table 3.

Table 3. Median parameter estimates (with 95% confidence interval), disaggregated by student proficiency

Bin	β_M	β_{MS}	β_{DS}
Low Proficiency	0.275 ± 0.162	0.102 ± 0.232	0.184 ± 0.315
Medium Proficiency	0.245 ± 0.083	-0.044 ± 0.133	0.655 ± 0.197
High Proficiency	0.002 ± 0.088	0.189 ± 0.208	0.947 ± 0.431

Do the relative values of the different types of practice follow the same pattern within each bin as the general estimates in Table 2? Mostly yes. The effect of MS stays trivial, and all the β_{DS}'s are larger than 0. Yet in some bins there are discrepancies from the general pattern. In particular, for the high proficiency students, the impact of massed practice on a word is not different from 0, significantly less than the impact of the same practice for the other two lower proficiency groups ($p < 0.05$). This trend of decreasing benefits from massed practice for more advanced readers is also reported in [5]. At the same time, for the same group, we find no significant difference between β_{DS} and the base line β_D ($\beta_D = 1$). The high values of β_{DS} indicates that proficient readers benefit almost as much from practice on similar words as they do from the word itself, which suggests that they may be sensitive to the morphemic structure of a word.

How do the transfer effects change as proficiency increases? Examining the column for β_{DS} values gives a suggestive picture. The effectiveness of distributed exposures to similar words increases from low proficiency to high proficiency students. Bootstrapping tests show that β_{DS} of high proficiency students is significantly larger than that of low proficiency students. The result supports our initial hypothesis that transfer increases with the growth of students' reading skills.

3.3 Greater learning gains, more transfer?

This analysis is similar to that in section 0. But instead of grouping students by their proficiency, we disaggregate the students by their improvement over the school year. We quantify students' improvement as the difference between their pretest and posttest grade-equivalent scores on the WI test. Again, the students are grouped into 3 bins: low gain, medium gain, and high gain. WI scores of low gain students increase no more than 0.6 from pretest to posttest. High gain students improve by more than 1.1 in WI score. Medium gain students lie in between. Table 4 gives the medians and 95% confidence intervals of the parameter estimates for each bin.

Table 4. Median parameter estimates (with 95% Confidence Interval) disaggregated by student gain

Bin	β_M	β_{MS}	β_{DS}
Low Gain	0.214 ± 0.193	0.076 ± 0.181	0.476 ± 0.218
Medium Gain	0.254 ± 0.165	0.086 ± 0.171	0.637 ± 0.212
High Gain	0.134 ± 0.081	0.056 ± 0.190	0.549 ± 0.211

Contrary to our hypothesis, none of the differences between bins turn out to be significant. This null result is a bit surprising and disappointing. The degree of transfer from similar words should be closely correlated with the students' awareness of the morphemic structure of words. In many previous studies, it has been shown that morphological awareness plays an important role in both vocabulary acquisition [10] and long-term reading development [11]. Why doesn't this trend show up in our analysis? There could be many reasons. For one thing, we are using the noisy test score difference to measure students' improvement. It is quite possible that a student did particularly well or poorly in one test, so test scores are imperfect reflections of students' ability. When we use only one test score, as in section 0, such noise should be washed out by aggregating over many students. However, when we use pre- to posttest gain as the measure, the variance in the two test scores combines, which results in even more noise: the difference between two noisy estimates is noisier.

3.4 Which student characteristics predict transfer?

In both section 3.2 and section 3.3, our studies are driven by our prior expectations. In this study, we adopt a different, bottom up approach to study which factors could affect the amount of transfer. Specifically, we build a linear regression model with various properties of a student as independents, to predict the estimated DS parameter for the student. We use DS as the dependent rather than MS, because the transfer effect from practice of type DS is significantly larger.

However, as mentioned before, individual estimates are extremely unreliable. In the case of DS, the standard deviation of the 328 students' estimates is 1600, which makes it unreasonable to model the exact value of DS directly. To get around this problem, we model the *rank* of DS for a student instead. Since there are 328 sets of parameter estimates, our dependent variable ranges from 1 to 328. We use three predictors. One is the students' grade (1—5). Another is the student's *grade-relative* WI pretest score, computed as WI score minus the student's grade, to eliminate the strong correlation between grade and WI test score. We use both grade and grade-relative WI, instead of just grade-equivalent WI. We treat grade-relative WI as a proxy for the student's relative proficiency at his or her grade level, while grade adds extra information about the student's experience in reading. The third predictor is the student's pretest score (20—80 in our dataset) on the ERAS measure of reading motivation [12].

The resulting model is barely significant ($p \approx 0.05$), with an R^2 of only 0.024. Among the three independents, grade is most predictive and reliable ($p \leq 0.05$), and grade relative WI score is borderline reliable ($p < 0.1$). Coefficients for all the predictors are positive, showing that transfer tends to increase with higher grade, higher WI test score, and greater interest in reading.

Considering that individual DS estimates are very noisy, the small R^2 is not surprising. The problem is that a lot of the variance is caused by random noise, which is essentially unexplainable, so R^2 is not a viable measure here. Therefore we take a different approach to evaluating our model, by looking at whether it separates the students into interesting groups.

Accordingly, we split all the students to 2 bins based on the standardized prediction of the model, which ranges from -1 to +1. We rank DS in increasing order, so the higher the rank, the larger the DS. Students in bin 0 (low transfer) have a standardized DS rank of less than 0. The rest of the students are assigned to bin 1 (high transfer). We adopt this 50-50 split because it's a reasonable and natural starting place in the absence of more information about the dividing point between low and high transfer. Then we compare the median of estimates for other parameters of the two groups of students, as shown in Table 4. For comparison, we also bin students into two groups by their original DS estimates, and compute the medians of parameter estimates for each group, given in Table 5.

Table 4. Comparison of median parameter estimates for two bins based on model prediction

Bin	L	A	B	M	H	HS	MS	DS
Low Transfer	0.132	1.067	-0.045	0.265	-1.311	-0.528	-0.004	0.402
High Transfer	0.136	0.464	-0.136	0.069	-1.232	-1.003	0.116	0.875
Sig. of Difference	0.2165	<0.0001	<0.0001	0.0165	0.4486	0.0044	0.245	<0.0001

Table 5. Comparison of median parameter estimates for bins based on original DS estimates

Bin	L	A	B	M	H	HS	MS	DS
Low Transfer	0.129	0.894	-0.072	0.163	-1.454	-0.500	0.282	-0.211
High Transfer	0.137	0.720	-0.091	0.184	-1.164	-1.590	-0.214	1.460
Sig. of Difference	0.066	0.0105	0.1179	0.3789	0.1455	0.0001	<0.0001	<0.0001

The result shows that the model is doing a decent job at separating the data, even better than the original DS does. The contrasts between the two bins in Table 4 are sharp (more so than in Table 5): students predicted to have low transfer have slower initial reading time (A = 1.07 vs. 0.464), learn slower (B = -0.045 vs. -0.136), benefit more from massed practice (M = .265 vs. .069), and transfer less (DS =.402 vs. .875). Moreover, all these differences are statistically significant. Therefore even though the model's R^2 is low and barely significant, the model separates DS much better than we expected. The clearer separation may be because our model smoothes out some random noise in the original DS estimates.

4. Limitations and Future Work

Though our learning decomposition model has refined the trial counts in the exponential model, it's still much simplified with respect to the actual complex cognitive process of learning. For example, the model assumes that the parameters are static for a specific student, while in reality the learning rate as well as the relative effectiveness of different types of practice will change over time with the development of the students' learning strategies. By simply partitioning all the practice on a word into non-overlapping types, our model also overlooks the interaction between different types of practice, such as order effects. We could introduce more complex terms into the model, but at the risk of reducing the data for each term and further increasing the variance in parameter estimates. As computationally efficient as it seems now (it takes only about 10 minutes to fit the nearly one million outcomes), the training process guarantees only a local optimum, and there is usually large error in parameter estimation for each individual model (though the median point is relatively stable compared to individual estimates). The model fit is rather low, only about 11.2% on average, similar to the number reported in the original learning decomposition model [5], hence not dismaying. Though we could argue that fitting individual data points tends to result in a poorer model fit than fitting a curve to aggregate performance, it's still a priority to improve the fit and robustness of the model.

The fact that our data comes from noisy ASR output also makes the estimates error-prone. In a test on data collected from the 2001-2002 school year, the ASR detected only about 25% of students' misreadings, and misclassified 4% of correctly read words as misread [13]. The best thing we could hope for, is that the ASR errors are random, and their effect on our measure of reading time would be diminished by aggregating over individual estimates so as not to blur the overall pattern.

Our future work will focus on improving the accuracy of the model, perhaps by modifying the form of the model to better fit the data. One possible improvement we are considering is to add another term that accounts for a student's total text exposures outside the Reading Tutor. Then we'll have a more accurate count of the student's prior word exposures, by adding this outside-tutor exposure estimate to the original within-tutor count. Another future direction is to extend our current study to other types of transfer. For example, we are analyzing transfer between rhyming words (i.e., words that share the same rime, such as "cat" and "hat"). We may also take into account the context effect (e.g., seeing a word in a novel vs. familiar context) in our model. For instance, we could use our approach to analyze whether it is better to read "the cat in the hat" twice or see "cat" the second time in some other sentence.

5. Conclusion

The contributions of this paper lie both in its methodology and in its results. There have been many studies in transfer of learning [14, 15], the paradigm of which is to carry out carefully designed controlled experiments. The method we describe is a simple yet potentially powerful alternative, and could be applied to many other intelligent tutoring systems. Our model exploits the massive and detailed (albeit noisy) machine observations of students' activities to make quantitative inference about their cognitive process. It is easily adaptable to changes in our underlying assumptions about the form of transfer. For example, we could look into transfer from rhyming words simply by making minor changes to Equation 3 and retraining the model. Moreover, so far as we know, previous studies of transfer at the individual word level [e.g., 14] looked only at transfer to previously unseen words. In contrast, our large data set lets us look at transfer to a word across multiple instances.

We also get some interesting and plausible results regarding transfer from practicing a word to learning similar words. We find that there is significant transfer from distributed practice on a word to similar words, but the effect becomes negligible when the student has already seen a word with the same root that day. Our hypothesis about the positive relationship between reading proficiency and the ability to learn from word roots is also supported by the result. Such findings could assist us in designing or evaluating fluency practice texts by predicting their effects for students at different levels of proficiency. For example, our result suggests that proficient readers learn significantly more than less proficient readers learn about a particular word from exposures to similar words. So for a less proficient reader learning to read a word fluently by practicing it on multiple occasions, it's important to practice the word itself each time, whereas for more proficient readers, it's almost as good to practice different words with the same root. Also, the less proficient reader derives some benefit from seeing a word repeated on the same day, though not as much as from seeing it on different days. In contrast, more proficient readers gain nothing from same-day repetitions, so text designed to improve their fluency should minimize unnecessary word repetition.

Another conclusion is that learning decomposition is a useful technique for investigating students' representation of knowledge. In particular, it provides a way to model transfer among similar skills without assuming a shared more general underlying skill, as previous approaches do [3, 4]. Our particular example here assumes such a skill (the shared root), but the model doesn't have to. For instance, we could have defined "similar" as "90% identical," or as "same except for 1 letter" (or *k* letters). This type of similarity doesn't correspond to sharing a single more general skill. Thus a model with explicit transfer has more (or different) expressive power than one based on shared abstract skills, thereby enabling us to gain more insight into a hidden cognitive process such as transfer of reading skill.

References

1. Gough, P.B. The Beginning of Decoding. *Reading and Writing: An Interdisciplinary Journal*, 1993. *5*(2): p. 181-192.
2. Mostow, J., G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M.B. Sklar, and B. Tobin. Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 2003. *29*(1): p. 61-117.
3. Chang, K.-m., J.E. Beck, J. Mostow, and A. Corbett. Using speech recognition to evaluate two student models for a reading tutor. *Proceedings of the AIED 05 Workshop on Student Modeling for Language Tutors, 12th International Conference on Artificial Intelligence in Education*, 12-21. 2005. Amsterdam.
4. Koedinger, K.R. and S. Mathan. Distinguishing Qualitatively Different Kinds of Learning Using Log Files and Learning Curves. *ITS2004 Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes*, 39-46. 2004. Maceio, Brazil.
5. Beck, J.E. Using learning decomposition to analyze student fluency development. *ITS2006 Educational Data Mining Workshop* 2006. Jhongli, Taiwan.
6. Porter, M.F. An algorithm for suffix stripping. *Program*, 1980. *14*(3): p. 130-137.
7. Beck, J.E. Does learner control affect learning? *Proceedings of the 13th International Conference on Artificial Intelligence in Education* 2007. Los Angeles, CA.
8. Cohen, P.R. *Empirical Methods for Artificial Intelligence*. 1995, Cambridge, Massachusetts: MIT Press. 405.
9. Woodcock, R.W. *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.
10. McBride-Chang, C., R.K. Wagner, A. Muse, B.W.-Y. Chow, and H. Shu. The role of morphological awareness in children's vocabulary acquisition in English. *Applied Psycholinguistics*, 2005. *26*: p. 415-435.
11. Deacon, S.H. and J.R. Kirby. Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics*, 2004. *25*: p. 223-238.
12. McKenna, M.C., D.J. Kear, and R.A. Ellsworth. Children's attitudes toward reading: a national survey. *Reading Research Quarterly*, 1995. *30*: p. 934-956.
13. Banerjee, S., J.E. Beck, and J. Mostow. Evaluating the Effect of Predicting Oral Reading Miscues. *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 3165-3168. 2003. Geneva, Switzerland.
14. Benson, N.J., M.W. Lovett, and C.L. Kroeber. Training and Transfer-of-Learning Effects in Disabled and Normal Readers: Evidence of Specific Deficits. *Journal of Experimental Child Psychology*, 1997. *64*(3): p. 343-366.
15. Taguchi, E. and G.J. Gorsuch. Transfer Effects of Repeated EFL Reading on Reading New Passages: A Preliminary Investigation. *Reading in a Foreign Language*, 2002. *14*(1).