Int J Artif Intell Educ (2015) 25:346–379
DOI 10.1007/s40593-015-0039-y

ARTICLE

A Measurement Model of Microgenetic Transfer for Improving Instructional Outcomes

Philip I. Pavlik Jr · Michael Yudelson · Kenneth R. Koedinger

Published online: 14 March 2015

Abstract Efforts to improve instructional task design often make reference to the mental structures, such as "schemas" (e.g., Gick & Holyoak, 1983) or "identical elements" (Thorndike & Woodworth, 1901), that are common to both the instructional and target tasks. This component based (e.g., Singley & Anderson, 1989) approach has been employed in psychometrics (Tatsuoka, 1983), cognitive science (Koedinger & MacLaren, 2002), and most recently in educational data mining (Cen, Koedinger, & Junker, 2006). A typical assumption of these theory based models is that an itemization of "knowledge components" shared between tasks is sufficient to predict transfer between these tasks. In this paper we step back from these more cognitive theory based models of transfer and suggest a psychometric measurement model that removes most cognitive assumptions, thus allowing us to understand the data without the bias of a theory of transfer or domain knowledge. The goal of this work is to help provide a methodology that allows researchers to analyse complex data without the theoretical assumptions clearly part of other methods. Our experimentally controlled examples illustrate the non-intuitive nature of some transfer situations which motivates the necessity of the unbiased analysis that our model provides. We explain how to use this Contextual Performance Factors Analysis (CPFA) model to measure learning progress of related skills at a fine granularity. This CPFA analysis then allows us to answer questions regarding the best order of practice for related skills and the appropriate amount of repetition depending on whether students are succeeding or failing with each individual practice problem. We conclude by describing how the model

P. I. Pavlik Jr (✉)
Institute for Intelligent Systems and Psychology, University of Memphis, Memphis, TN, USA
e-mail: ppavlik@memphis.edu

K. R. Koedinger
Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: koedinger@cmu.edu

M. Yudelson
Carnegie Learning, Inc., Pittsburgh, PA, USA
e-mail: yudelson@cmu.edu

allows us to test theories, in which we discuss how well two different cognitive theories agree with the qualitative results of the model.

Keywords Human learning · Pedagogical strategy · Transfer of learning · Student modelling

Introduction

When considering the effects of education it is plausible to argue that transfer of learning to new situations is the main underlying goal of education. Because of this dependence of education on transfer, work by many prominent researchers has concentrated on transfer related phenomenon (e.g., Anderson and Fincham 1994; Bransford and Schwartz 1999; Chen and Klahr 2008; Gentner et al. 2003; Gibson 1940; Gick and Holyoak 1980; Postman et al. 1968; Rickard and Bourne 1996; Rittle-Johnson et al. 2008; Sloutsky et al. 2005; Son and Goldstone 2009; Sternberg 2008). Transfer of knowledge learned during prior events to a new performance context can be challenging for students, so better measurement of transfer that furthers our ability to configure instruction to help students learn in ways that are transferable is paramount to the goal of effective education. A long-standing open question is what is the nature of the mental capacity that "carries" or transfers what is learned in the source instructional context to the target performance context? Is it general faculties of the mind (Singley and Anderson 1989), identical elements represented as stimulus–response bonds (Thorndike and Woodworth 1901), more general elements represented as production rules (Singley and Anderson 1989), or hierarchies of relational structures (Falkenhainer et al. 1989; Hummel and Holyoak 2003). This question is of critical importance to understanding human learning and engineering better educational systems, but it also occludes the often more immediate problem of how an instructional designer should diagnose and improve specific classes of learning objects (e.g., problem types) to maximize their effects.

Toward that end, a simplified approach used frequently in the AIED community is to identifying the carriers of transfer by enumerating the relevant components of knowledge (cf., Koedinger et al. 2012) and map them to the tasks in source instruction and target performance. This approach has been employed in psychometrics (e.g., Tatsuoka 1983), cognitive science (e.g., Koedinger and MacLaren 2002), and most recently in educational data mining (e.g., Cen, et al. 2006). These newer psychometric models of transfer, based on a question by skill matrix or "Q-matrix" that maps skills to tasks, have proven useful to handle the massive and growing data streams coming from the increasing use of educational software in the classroom and for homework. The Q-matrix method is applied by assigning some number of skills to some number of item classes, according to what skills are needed to do the performances for that class of items. These latent skills or "knowledge components" are often assumed to be many fewer than the number of problem/item classes, so the Q-matrix is thought to be useful in representing and tracking learning in a domain.

These models come from a growing field, educational data mining, which has been developing methods to detect and summarize the meaning of educational data to maximize its value to the educational research community (Romero and Ventura

2007). While educational data mining has many methods, this paper focuses on model-based discovery, a technique that uses mathematical models to create the summary understandings that can then feed back into improvements in educational technology, and hopefully education more generally. Model-based discovery is a new area of educational data mining, and publications stress the importance of these methods (e.g., Baker and Yacef 2009). More generally, educational data mining fuses the concerns of several of the cognitive science related disciplines including educational psychology, artificial intelligence, psychometrics, and cognitive psychology.

Our goal in this paper is to demonstrate the CPFA (Contextual Performance Factors Analysis) model, a data-driven approach that provides a rich starting point in the process of learning theory verification and in the development of better learning objects. CPFA provides a numerical summary of data that would otherwise be difficult to process, by starting with simple statistical assumptions and different classes of learning objects. In specific, this paper addresses the question of how this approach can be used to understand learning and transfer data from pre-algebra learning objects. We employ model-based discovery to pinpoint instructional activities (i.e., contexts of learning) that are unproductive for future performance. A key point about the results of such CPFA models is that they can summarize complex patterns in a dataset within a small table of numbers. In the discussion and conclusion, we show how this interpretability leads to design decision making about the learning objects measured. We also demonstrate the model's flexibility by showing how it captures situations not represented by the more limited 'Q-matrix' formalism for representing transfer components. Finally, we describe how analysing the patterns of parameters in a model facilitates comparison of alternative theories about the extent of transfer across different kinds of activities.

A distinctive feature of CPFA is that it models both asymmetric transfer and negative transfer, unlike past Q matrix based approaches. Asymmetric transfer describes the situation where practice of item A transfers to performance on item B, but where similar practice on item B does not benefit performance on item A. Such asymmetries in transfer are not uncommon, and specific results in the literature suggest a powerful advantage for the transfer of general math abilities into more specialized applications of math. For example, in Bassok and Holyoak (1989), there are advantages for the students that practice problems in a math domain and transfer to physics problems as compared to students who practice the same physics problems first and show less transfer to the same math problems. CPFA also models negative transfer effects, where practices of an item can actually reduce performance on later items, as might be caused if one item type resulted in misconceptions about how to solve the other item type.

Our intuition about how to model such situations is similar to the resolution provided by Singley and Anderson (1985), looking at how general and specific skill generalized to different contexts of application. In their case they were looking at how different amounts of practice transfers between the learning of different text editors. They found that using 2 types of knowledge component (general and specific) it was possible to predict the transfer results using a function that weighted the importance of general and specific practice differently. So for instance, with the EDT editor, general learning was almost 3 times as strong as specific learning, while for the EMACS editor general learning was less strong, and specific learning was proportionally more important. This model is similar to ours because unlike standard Q-matrix models we are also

proposing that the amount of transfer is contextually dependent. In contrast, Q-matrix models typically assume that knowledge components contribute to performance independent of their context, in a discrete, all-or-nothing fashion. Unfortunately, the idea that knowledge transfer depends on context is complex and it is not trivial to consider how to use Singley and Anderson's (1985) ideas to create a statistical model to generally detect when transfer depends on context.

This paper summarizes our efforts toward a psychometric model of transfer that makes no strong assumptions about the mechanism of transfer (it is not a theory of the cognitive structures or processes), but rather measures the transfer between classes/types of items after controlling for the random effects of item and subject and the fixed effects of different item classes. We hope to facilitate learning scientists in measuring and identifying which instructional activities lead to greater transfer of learning and in understanding why. Because CPFA is independent of any particular theory of the process of transfer, it can be applied more broadly and be used to quantitatively summarize the key results that a process-based theory of transfer must satisfy. Because this model is so flexible, it will tend to capture effects that many theory-based models may exclude from consideration. We illustrate this flexibility in contrast to the Q-matrix approach, which conforms with theories that are limited to only symmetric and positive transfer. Similarly, we describe how the model can be used to support hypothesized qualitative relationships (e.g., math problems transfer better to physics than the other way around) by an analysis of whether the parameters in the model conform to the patterns implied by the theory.

Pre-algebra was selected as a domain for this transfer research because it has good affordances for experimentation to understand transfer and is an area in which real students might benefit greatly from improved instruction. Pre-algebra has excellent affordances for research for a few reasons. First, math generally has a componential nature that makes it easier to identify why and how problems are related in terms of their logically necessary solution procedures. Second, math problems can often be made so that they are short and single step (or at least have a single unambiguous answer), but nonetheless are conceptual since they involve categories of examples having a coherent solution schema where people can form consistent misconceptions. Third, math is rich with different problem contexts or situations (e.g., story problems and number lines) and so allows us to understand how representational features and relationships can block or enhance transfer of learning to future performance.

The data we analyse here were produced from two pre-algebra "warm-up lessons" used within the context of the Carnegie Learning Inc. pre-algebra computerized intelligent tutoring system (called Bridge to Algebra; http://www.carnegielearning.com/). For each student, each warm-up lesson consisted of 16 single step problems selected randomly without replacement from a set of 24 possible items. The 24 items were themselves split into various related item-types, and our method investigates learning and transfer effects during each warm-up.

Contextual Factors Approach

Our method for analysing transfer from instructional events is called a contextual factors approach to highlight the notion that the interaction of the driving context of

learning and the target context of performance determines transfer performance (Pavlik Jr et al. 2011). The contextual factors approach uses categories that cross learning contexts by future performance contexts, where by "context" here we mean a set of similar tasks. These categories are used to determine what prior events need to be considered when predicting future performance. More specifically, if A and B are types of tasks (e.g., A might be word problems and B might be equations), we have four categories of learning transition A→A, B→B, A→B, and B→A. The last two categories are commonly called transfer to distinguish them from the first two, which can be thought of as simple learning. The contextual factors approach postulates that these pairwise learning and transfer categories capture the specific nature of the relationship between the type of practice (procedural or declarative), concepts and contexts underlying each task.

We label our most recent model within this approach "contextual performance factors analysis" (CPFA, elaborated below). CPFA fuses an interest in both the performance and transfer context. We contrast it with a simpler model called "performance factors analysis" (PFA, described below). CPFA uses mixed-effect logistic regression to estimate the strength of the effect of the contexts of practice and performance on the performance result (correct or incorrect) subsequently observed. Because CPFA involves fine grained event-by-event tracking of the developmental change, CPFA allows the researcher to take a "microgenetic approach" (Siegler and Crowley 1991). As described by Siegler and Crowley, microgenetic analysis 1) measures individuals throughout the period of change, 2) has a high observation rate relative to the rate of change, and 3) involves intensive trial-by-trial analysis to infer the process of change. In contrast, pre-post assessment designs to test interventions can show macro level change, but without trial-by-trial analysis of effects, the source of the change is typically more difficult to pinpoint.

An example helps clarify these ideas. In our first set of items, there were two kinds (or item-types) of least common multiple (LCM) items. A first item-type was a simple textual item that asked the student a question like, "What is the LCM of 3 and 5?" while the other item-type were story items, "Sally visits her grandfather every 3 days and Molly visits him every 5 days. If they are visiting him together today, in how many days will they visit together again?" The item-types were matched by using the same numeric content in the items for each item-type to counterbalance our design. The student responds to such questions by typing in an answer and system provides feedback. If the answer is correct (a "success"), the system gives correctness feedback (a checkmark). If the answer is incorrect (a "failure"), the system provides the correct response, and an icon indicating that the student should review the correct answer. Thus the only instruction students received was the presentation of the correct answer, from which they needed to infer how it answer was generated.

Given these two item-types, Table 1 shows the eight types of learning and transfer that CPFA tracks. The model separates effects given prior success and failures, since prior work shows that the benefit of learning is different from a success than it is from a failure (Pavlik Jr et al. 2009). The model also separates the relationship between the item-type for the source of learning or transfer and that for the target. When the source and target are the same, the Story→Story effects and the Word→Word effects, we refer to this as "learning" and we distinguish learning given prior successes and learning given prior failures. When the source and target are different, the Story→

Table 1 An example of the 8 possible learning and transfer relationships between two classes of items, "Story" and "Word"

Source item-type		Target item-type	
		Story	Word
Story			
	Success	learning given success	transfer given success
	Failure	learning given failure	transfer given failure
Word			
	Success	transfer given success	learning given success
	Failure	transfer given failure	learning given failure

Word effects and the Word→Story effects, we refer to this as "transfer" and we distinguish transfer given prior successes and transfer given prior failures.

History of the Method

Formally, the model we have created can be traced from the development of item response theory (IRT) (Rasch 1966). IRT was developed to understand and evaluate results from academic testing by allowing a modeler to fit parameters to characterize performance on a set of items by a set of students. As a form of logistic regression IRT predicts 0 or 1 (dichotomous) results like the results of individual items for students. In the case of a 1 item parameter IRT model, this result is predicted as a function of student ability minus item difficulty (x), which is scaled to a probability estimate by looking up the probability from the logistic function cumulative distribution, in which $p=(1/(1+e^{-x}))$.

IRT has had a long developmental history. For example, around the time others were exploring Bayesian models of students (R. C. Atkinson 1972), the Linear Logistic Test model (LLTM) was developed, which introduces the idea that multiple discrete "cognitive operations" may combine to determine an items difficulty (Fischer 1973). While IRT might be called a behavioral description, since it does not have parameters mapping to mental constructs, LLTM goes a step further by proposing a set of knowledge components that combine to produce an item's difficulty. This work maps closely to work that describes "knowledge spaces" or what have come to be known as Q-matrices, which is a means to lay out a tabular matrix to specify exactly what knowledge components, misconceptions or cognitive operations are needed for each practice or test item (Barnes 2005; Tatsuoka 1983).

But the LLTM model alone does not capture learning. On a path toward modeling learning, we trace the development to work by Scheiblechner (1972) which uses the LLTM model, but also examines changes in difficulty as people repeat knowledge components. This work is well reviewed by Spada and McGaw (1985) who unpack the history of these sorts of models, which have persisted (e.g., Draney et al. 1995) and come to be known as the additive factors model (AFM, Cen et al. 2008) within the AIED/EDM communities. AFM is relatively straightforward, and proposes that we should add a term to these IRT type models that captures the prior quantity of practice

for each knowledge component as a linear effect. These skill tracking models have been combined with the Q-matrix knowledge component models in part perhaps because while formally identical to the psychometric analysis in LLTM, Tatsuoka's (Tatsuoka) work more clearly explains this method from a less mathematical/ more pedagogical way that has appealed to learning science researchers. A variant of this LLTM/AFM model is built into the DataShop repository at Carnegie Mellon University and is currently being used to do automated search for better Q-matrixes to represent the skills in different educational systems (Stamper and Koedinger 2011).

It was into this arena that performance factors analysis (PFA) was introduced as a way to improve the AFM model still further by noting the importance of using the prior successes and failures as a powerful source of information for predicting future performance. This change improves model fit considerably, making logistic regression at least as accurate as Bayesian knowledge tracing (BKT) a Markov model that is used to fit similar data (Corbett and Anderson 1992; Gong et al. 2010; Pavlik Jr. et al. 2009). The PFA model assumes that there are two fundamental categories of practice, success and failure. As the psychological literature suggests, successes (in contrast to review after failing) may lead to more production-based learning and/or less forgetting (Carrier and Pashler 1992; Karpicke and Roediger 2008; Pavlik Jr. 2007; Thompson et al. 1978) and may lead to automaticity (Peterson 1965; Rickard 1997, 1999; Segalowitz and Segalowitz 1993). In contrast, failure reveals a need for declarative learning of problem structure and verbal rules for production compilation (Taatgen and Lee 2003). We might expect learning after failure to depend most importantly on the feedback that occurs after the task is complete.

While AFM and PFA have been forward developments, these models do not fit all patterns in the data that may occur in nature (as we will show) and are based on assumptions about the Q-matrixes (assumptions of the LLTM method) that, unless satisfied, make the entire procedure subject to error. These problems may make it difficult to use the result of the AFM and PFA analysis to improve instruction. These problems with standard Q-matrices include begin unable to capture effects such as asymmetric learning, negative transfer, and conjunctive skill situations that do not involve the compensatory trade-off implied when we add the different concepts strengths as the linear logistic model stipulates. For these cases we have been working on the contextual factors approach, which looks for direct transfer effects between empirically constrained problem-type equivalence classes.

As stated earlier, in its simplest form, CFA does not include the tracking of success and failure, but merely tracks the effects of item classes on other item classes. While this is useful by itself to see whether transfer is asymmetric, bidirectional, etc., CFA can also be combined with tracking of success and failure in the model to create CPFA, which tracks the 8 categories of prior practice results as described in the prior section. Further, unlike methods that are based on a Q-matrix – the sparse structure assigning latent skills to items or tutor steps, which then transfer symmetrically – (e.g., AFM and PFA), CPFA explicitly assumes asymmetric (one-way) transfer may occur and does not presume any latent factors exist (Pavlik Jr et al. 2011). Rather, in CPFA each item-type causes different things to be learned, which may transfer either to future practice with the same item category (learning effects), or future practice with items in other categories (transfer effects, e.g., LCM items without stories→LCM items with stories). While each of the transfer effects is somewhat like learning a skill in the knowledge

space theory (Tatsuoka 1983), the model does not assume that any learning or transfer necessarily generalizes to other item-types. While asymmetric transfer is not always the case, there are cases where it occurs strongly (e.g., Anderson and Fincham 1994; Bassok and Holyoak 1989) and it is useful to have a method to detect it.

In the simplest case CPFA can be used to analyse learning and transfer between 2 item-types A and B given randomly (with replacement) for 2 trials. Since the 2 trials are given in a random order with replacement, the possible item orders of the 2 trials are AA, AB, BA, and AA. Typically, in this sort of paradigm we assume each repetition of an item-type is drawn from a population of items, so that while AA means 2 repetitions from the same item-type, they are not verbatim identical repetitions, since there may be variability within a population. In this simplest case we have 4 learning contexts (A or B)×(success or failure) and 2 performance contexts (A or B). CPFA says that we cross these categories to get the full transfer relationship matrix which specifies the relationships in Table 2. This matrix lists how each of the eight situations is captured by a different model coefficient (coefficient γ's are for success and ρ's are for failures). Having this example of an item-type A and an item-type B allows a more specific explanation of the how source contexts influence subsequent target contexts in CPFA. *In CPFA there is a single coefficient capturing each possible sourceXoutcomeXtarget combination*. Therefore, because there are four categories of prior *sourceXoutcome* practice – success on A, success on B, failure on A and failure on B – and because there are two categories of future *target* practice – A or B, the model has eight learning and transfer parameters (four conditions of prior practice by two conditions of future practice).

This model of the eight relationships is assumed to be additive, so multiple instances of possibly different source item-types are added to determine the performance for a target item-type. Such an additive model is easy to consider as a linear equation where the performance is predicted by summing the number of practice events times the effect of each event. However, since our predicted quantity is a probability, we model these additive effects as a logistic regression, since logistic regression is the standard model to convert a linear equation result to a probability. Table 3 shows the model's linear equation (Eq. 1) and logistic conversion function (Eq. 2). In addition to the 8 coefficients for the 8 predictive relationships in Table 2, the mixed-effect model also includes fixed-effect parameters that captured the difficulty of each item type, and random-effect distributions for student and item, capturing prior knowledge by student and individual item difficulty level. Ultimately, due to the mixed-effect model form, the model is well suited to including other hierarchical factors of the learning, such as the learning context (e.g., working at home, working at school), experimental conditions, and

Table 2 Learning and transfer effects in the model (coefficient γ's are for success and ρ's are for failures)

	Success	Failure
A→A (learning)	γ_{AA}	ρ_{AA}
B→A (transfer)	γ_{BA}	ρ_{BA}
B→B (learning)	γ_{BB}	ρ_{BB}
A→B (transfer)	γ_{AB}	ρ_{AB}

Table 3 Mixed-effect logistic regression equation

$$m_i = \beta_0 + \beta_A + \gamma_{AA}S_{AAi} + \gamma_{AB}S_{ABi} + \gamma_{BA}S_{BAi} + \gamma_{BB}S_{BBi} + \rho_{AA}F_{AAi} + \rho_{AB}F_{ABi} + \rho_{BA}F_{BAi} + \rho_{BB}F_{BBi} + u_i + e_k, u_i \sim N(0, \sigma_u^2), e_k \sim N(0, \sigma_e^2)$$ Equation 1

where:
β_0, is an overall fixed intercept,
β_A, is an intercept for the item-type contrast level,
γ_{**}, effects of counts of prior successes (S_{**i}) relevant to the item (across students),
ρ_{**}, effects of counts of prior failures (F_{**i}) relevant to the item (across students),
u_i, random effect intercept for each student i,
e_k, random effect intercept for each item j

$$p_i = \frac{1}{1+e^{-m_i}}$$ Equation 2

random school or classroom effects, but in this paper we have kept the scope limited to item-types and performance categories.

See Eq. 1, Table 3. Our model includes fixed effect intercepts, which capture the average difficulty of the 2 item-types. β_0 is the intercept for a 1st type of items, and β_A provides the difference for the 2nd type. The model also includes random effect intercepts, characterizing the prior knowledge of individual students and the prior difficulty of individual items. The use of the random effects encourages the gamma and rho parameters for the effect of successes and failures to track changes in performance due to learning and transfer, rather than track student or item differences. If we did not use these random effects, the model γ and ρ parameters would tend to find values that mostly track prior learning and transfer rather than changes to prior learning. In other words, if we do not account for prior knowledge, the models estimates of learning and transfer will be confounded. For this reason it essential to use these random student intercepts if we hope to trust the transfer implications of the model.

We fit these models with the lmer function in the lme4 package from the R Project for Statistical Computing. As a reviewer of this paper noted, a model distinguishing the effects of better prior knowledge from the effects of current learning is desirable, but it is in practice very difficult to separately fit parameters characterizing both the *change in the estimate of prior probability success* and the *learning effect of practice* since they cannot be measured independently. Because of this, it is important to note that the model we have settled on emerged from intensive incremental model search accompanied by cross-validation tests. For each of the models a 5-fold cross validation over 20 randomized repetitions was computed. The computation was repeated twice. The first time, cross-validation was user-stratified, namely 80 % of user data was used for training and 20 % for testing. The second time, cross-validation was item stratified. For the CPFA model we have described, the signs and magnitudes of model parameters we discovered during model fitting were confirmed during cross-validation. Specifically, our CPFA model cross-validated with a correlation of .567, despite the fact that in validation folds we did not use any subject parameters, indicating that the .567 correlation came from the skill model alone. We have omitted these model validation details from this paper due to their length and complexity, but some detailed examples are given in prior work (Yudelson et al. 2011).

This model validation provided us assurance that our models were not just finding idiosyncratic patterns in the data; rather they are finding patterns that generalize to

unseen data. This validation helps answer a reviewer of the paper who was concerned that our success and failure parameters might be confounded with the student ability. While the reviewer was correct in principal, the cross-validation showed that using random effects models compensates for the confounding of the success/failure parameters and the student parameters. Random effects have this effect because they prevent the over-fitting of the subject variance that occurs with fixed effect student parameters (overfitting that results in success and failure parameters that are degenerate) and models with no subject effects (where under fitting results in success and failure parameters that track subject ability rather than learning) (as shown in Yudelson, et al. 2011).

Additionally, in the results below we report the simplified CFA model without this confound, to further confirm that the overall transfer patterns match the averaged transfer patterns in CPFA, thus establishing that the success and failure counting was not disrupting the overall model structure. In sum however, it is hard to see how this issue seriously troubles our desire to use the model for instructional optimization, since, assuming most of the learning parameters in a particular model of data are positive, showing that the model captures the overall positive slope of learning, negative coefficients must either indicate that these categories of performance are especially sensitive to student variability and tend to identify low prior knowledge or ability students OR that the item causes poor learning. In either case there is a clear need to improve the instruction in these practice contexts with near 0 or negative learning coefficients, either to scaffold the low knowledge/ability students or to make the item more instructional effective generally.

Methods

We gathered our data from the classroom but used randomization of item selection and sequence for each student because we were concerned about problems with modeling data containing various sources of bias (Shadish and Cook 2009). Often this bias in educational data occurs as a "confound" caused by the adaptive decisions. For example, in some systems using mastery based practice (Corbett and Anderson 1992), good students are more quickly skipped to the next unit or concept, which means that more data is collected from the poorest students for any particular unit. Even more serious data bias in learning curves can be caused if item order is not randomized. Similarly, systems that collect data while adjusting item difficulty to scaffold individual students tend to bias data by providing more easy items for lower performing students. These sorts of biases also affect inferences about transfer, for example if the practice order is always A→B it is impossible to compare the symmetry of transfer in the data.

Such data collection biases may limit causal inference from educational data mining results. To alleviate such limitations, we used a classroom experimental method that, by randomizing the order of problems more like a psychological experiment, allows unconfounded analysis of sequences of practice from a set of pre-algebra problems using a model taken from educational data mining. In general, methods to get more naturalistic educational data combined with rigorous control have been called "in vivo experimentation" for the way that they blend experimental method with an attention

to the real life issues of classroom learning (Koedinger et al. 2009; Koedinger and Corbett 2010), such as student attendance or the distractions that a classroom presents compared to the lab.

Setting

Data on transfer was collected from a Florida charter school both from classroom work on a computerized educational tutoring software program and from homework on the same system. This natural setting varied between individuals, but because the study used full random assignment of students to condition and items to student, the data can be used for post facto analysis of putatively causal effects. While our ten sets of intervention items were placed as part of the Bridge to Algebra product from Carnegie Learning Inc., we are not examining the Carnegie Learning system, but rather using it as a vehicle to deliver our intervention. Nevertheless, each of our intervention units did fit in the curriculum sequence in the Carnegie Learning system, so our interventions were appropriate for each student's current progress in the Carnegie Learning system.

Population

Approximately 250 6th and 7th graders participated (ages ranged from approximately 10–12 years). Participating classes included all levels at the school that used the Bridge to Algebra tutor. The exact count of participating students for each warm-up is listed in the results section.

Experimental Design

The experimental side of this project is best described as an experimental design in a naturalistic context, but this paper focuses on post-hoc model-based discovery methodology to analyse the implications of the student results. The research design used 10 sets of 24 individual pre-algebra single step questions on a variety of content. The 10 interventions we gave were split into item-types according to systematic analysis of their features. For each of the 10 interventions students were each quizzed on 16 randomly selected items from these sets of 24 possible items. We analyse 2 of the 10 warm-ups for this paper. These 2 warm-ups allow 3 examples of the method, since the first warm-up is analysed for 2 different item-type splits.

The students were randomized (by person) into 1 of 4 conditions for each warm-up. This paper does not report on the condition effects, but we present the conditions for completeness. In the standard condition, students were quizzed on the items (with a 10 min threshold to respond). If they responded correctly, there was a .75 s interval where a correctness indicator (a check mark in green) appeared, and then the next problem began. If they responded incorrectly, they were presented the correct response for 18 s. In the direct instruction condition, the trials occurred just as in the standard condition for problems 1–4 and 13–16, but during problems 5–12 a hint was presented on the screen at the same time as the problem, and also presented during the feedback with the answer to the problem. In the inference condition, once again problems 1–4 and 13–16 were presented as in the standard condition, but for problems 5–12, one of the problems was presented as a worked example, and student needed to fill in a

missing vocabulary word in a hint message text (cloze question). Both the worked example and full hint were given as feedback (note, in this condition we did not model trials 5–12, but rather modelled only the 8 trials with problems, assuming no effect of these inference trials in our model as indicated by our initial analysis of conditions). In the analogy condition, everything was the same as above for trials 1–4 and 13–16, but for trials 5–12 we presented a worked example item (from the left out 8 items in the set) for each of the trials in addition to the standard problem. This paper does not analyse the effects of the 4 between-subjects conditions for each unit, because prior analysis showed the main-effects of condition were non-significant.

Finally, students with 0 % correct or 100 % correct responding for the warm-ups were filtered from analysis, because it seemed they should not be allowed to bias the random effects subject distribution. In the case of 0 % performers, there was the possibility that they were merely noncompliant and may have done better if they engaged. In the case of 100 % performers, the possibility was that students could be using an outside resource for the warm-ups, since some of the work was done at home. In either case, there did not seem to be an advantage to retaining either group of students, since they merely shift means rather than show any variance as a function of learning and transfer conditions. We report the count of students in each case removed.

Results

Warm-Up 1 Example

Warm-up 1 contained items that addressed least common multiple (LCM) skills. These items presented the student with two numbers and asked them to produce the least common multiple. Problems were classified according to two factors. The first item-type factor distinguishes whether the problem could be solved correctly by simply multiplying the givens, for example, the least common multiple of 3 and 5 is 15. We called these "partial strategy" problems "Product" problems. Problems that cannot be correctly solved by multiplying the givens (e.g., least common multiple of 4 and 6 is 12 and not 24) will be referred to as "LCM" problems. The second item-type factor addressed how the text of the problem is presented. There were longer "Story" problems that place the problem in a context, and there were shorter "Word" problems that posed the question without a context. Examples of problems, corresponding factor labels, and number of problem items in each category are given in Table 4. While the spit between Story and Word was a planned contrast, the split between Product and LCM items was discovered in the analysis, and this accounts for the unequal number of problems per category (7 vs. 5 rather than 6 vs. 6). Before building the CPFA model we filtered the original warm-up 1 dataset of 3616 data points from 255 students down to 3520 data points from 247 students by removing students that demonstrated 0 % performance (1 user) and 100 % performance (7 students).

Product vs. LCM Item-Types Data and Model First, we examined Product and LCM item-types. Before describing the model fit details, we first describe learning and transfer patterns in the data as they are often described in transfer experiments (e.g., Gick and Holyoak 1983) where there are just two stages of observation. Figure 1a

Table 4 Item examples and corresponding factors for warm-up 1

Example item texts	Factor 1	Factor 2	No. of items
What is the least common multiple 4 and 5?	Word	Product	7
What is the least common multiple 4 and 6?	Word	LCM	5
Sally visits her grandfather every 4 days and Molly visits him every 5 days. If they are visiting him together today, in how many days will they visit together again?	Story	Product	7
Sally visits her grandfather every 4 days and Molly visits him every 6 days. If they are visiting him together today, in how many days will they visit together again?	Story	LCM	5

shows changes in students' average success rates from trial 1 to trial 2 for the item-types. At trial 1, two points are shown, each representing about half of the trial 1 data, namely average success rates for students getting a Product or LCM problem first. At trial 2, four points are shown, representing four different item-type combinations that students saw, namely, Product-Product, Product-LCM, LCM-Product, and LCM-LCM – each aggregating roughly one quarter of the trial 2 data. Solid lines connect attempts for identical item-types, denoting learning. Dashed lines connect attempts for different item-types and denote transfer. For example, for trial 2, the labels read "first type – second-type" so the notation LCM-Product, indicates performance on a Product item that was preceded by an LCM item. So, in Fig. 1a the top two lines show that students performed better on Product problems on trial 2 than they did on trial 1 (there is a learning gain) and more so, interestingly, when trial 1 practice was on an LCM item rather than a product item.

The purpose of Fig. 1a is to show learning and transfer as students start each new warm-up, since early gains in success probability are often greater and diminish as success approaches the ceiling of 100 % correct. However, at this grain-size (only averaging trial 1–2 performance), the values are noisier and as a result, the changes in

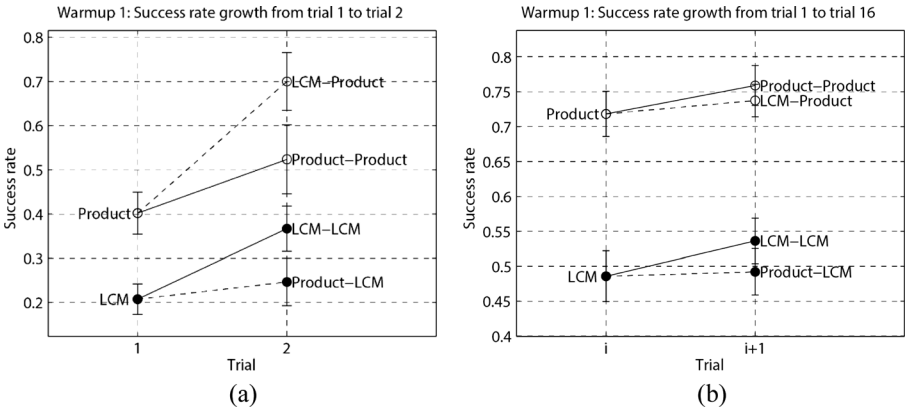

Fig. 1 Success rate growth for "Product" and "LCM" problems in warm-up 1: **a** from trial 1 to trial 2, **b** from trial i to trial i+1 averaged across all 16 trials. 1 *SE* error bars

success rates attributed to learning and transfer effects between trials 1 and 2 might not reflect general changes throughout practice. To get a more complete estimate of the relative differences between changes in success rates attributed to transfer and learning effects, we produced Fig. 1b. In Fig. 1b, the positions of six points are averaged across 15 trial pairs: 1 and 2, 2 and 3, 3 and 4, 15 and 16. Naturally, the positions of the points go up as the number of trials grows due to learning. However, in this particular case, we are more interested in the magnitudes of changes in the success rates – how far up or down the lines go. Note that this averaging of consecutive pairs blunts rather than exaggerates transfer effects because some AB sequences will occur in a context, such as BBAB, where the prior learning effects might overwhelm subsequent transfer.

As shown in Fig. 1a, LCM success rate increases from .21 to .37, Product success rate increases from .40 to .52 from trial 1 to 2. However, Product-to-LCM "transfer" success rate increase by merely .04 (.21 to .25), while the LCM-to-Product "transfer" success rate goes up by .30 (.40 to .70). The averaged view of learning and transfer in Fig. 1b is less pronounced, but the pattern is the same: learning is positive for both item-types, while transfer is asymmetrical. Table 5 presents a summary of the parameters for the Product vs. LCM CPFA model. Parameter values are on a logistic scale.

Product vs. LCM Item-Types Discussion Interpreting the logistic regression model requires an understanding of how the parameter differences actually translate to student change in percent correct (since odds ratios, the other alternative interpretation, are arguably less clear for this purpose). In other words, what does a .1 parameter value increment indicate in terms of percent correct? This can be exactly computed using Eq. 2, which implies that the effect of the parameter varies with the sigmoidal slope of the logistic cumulative distribution function. However, it is more useful to have a few concrete examples to understand the function quantitatively. So, we note that if probability is currently .5 (50 % correct) the .1 increment brings the learner to approximately 52.500 % probability of success. This would indicate 10 practices result in 25 % gain, except for the fact that the slope of the function decelerates as it approaches 0 or 1. Again, concretely we see this means that if probability is .75 there

Table 5 Summary of the Product vs. LCM item-type CPFA model parameters for warm-up 1 (including the full dataset shown in Fig. 1b)

Param.	val.	p-value†	Param.	val.	p-value†	Note
β_0	−.919	.000***				Overall intercept (LCM item-type)
β_A	1.250	.000***				Intercept modifier for Product item-type contrast level
γ_{AA}	.208	.000***	ρ_{AA}	.148	.040*	Learning from Product to Product
γ_{BA}	.094	.040*	ρ_{BA}	.083	.040*	Transfer from LCM to Product
γ_{BB}	.420	.000***	ρ_{BB}	.075	.030*	Learning from LCM to LCM
γ_{AB}	.020	.610	ρ_{AB}	−.013	.830	Transfer from Product to LCM
σ^2_u	.424					Variance of student random effect intercept
σ^2_e	.408					Variance of item random effect intercept

† Significance codes: . -p≤.1, *-p≤.05, **-p≤.01, ***-p≤.001

Success (γ) and failure (ρ)

is only a .0183 probability increase to 76.823 % from a .1 parameter change. This decrease in the increase continues, so at 95 % there is only a .00454 increase in probability to 95.454 %. The distribution is symmetric, so the same effect applies as values approach 0.

Using these guidelines helps us evaluate the strength of the learning and transfer that is captured by the model for this data set. We can see asymmetric transfer in the parameter estimates, which are statistically significant for success and failure transfer from LCM (.094 and .083, which implies a little less than 2.5 % transfer also given that starting probability is bit more than .5 i.e., -.919+1.250). We do not see this for success and failure transfer from Product (.020 and -.013). Again it is useful to note that these parameters represent the *cumulative* effects of the count of prior practices for the 4 possible events on each of the 2 contexts of application. The pairwise comparisons of the asymmetric transfer parameters (e.g., success transfer A→B, .020, compared to B→A, .094) were not quite significant by themselves. But we can collapse the success-failure distinction and examine asymmetry using a four-parameter model (for A→A, B→B, A→B and B→A trials) rather than an eight-parameter model. In this aggregate model (LCM→LCM=.237***, Product→Product=.196***, LCM→Product=.099** and Product→LCM=.021) we find that overall transfer from LCM to Product is significantly stronger than transfer from product to LCM (t=2.3, df=247, p<.05). Note that this simple contextual factors only model is extremely similar to the averages of the average of success and fail parameters in the full CPFA model. This shows how the addition of the performance tracking does not alter the overall transfer predictions, but rather provides additional useful information about which types of performance predict the need for improvements in instruction (either because the item does not cause learning or because the item identifies students that need remediation).

To summarize, we see that the product items simply do not provide transfer appropriate processing (Morris et al. 1977) for the LCM items. In contrast, the general LCM strategy subsumes the Product strategy and therefore transfer of learning does flow from LCM practice to product item performance. Similar asymmetric subsumption relationships have been found for English grammar therapy in agrammatic aphasics learning sentence structure with three or two argument verbs. While three argument verb training transferred to two argument verb performance, two argument verb training caused learning, but not transfer to three argument performance (Thompson et al. 1993).

In addition to asymmetry, we might also wonder if Product practice actually causes some student misconceptions. However, transfer-from-Product parameters did not indicate negative transfer overall (they are not significantly negative at .020 and -.013), just no transfer. Despite this, there were consistent misconceptions in students' incorrect responses. In 26 % of LCM error trials, a product of the two numbers was produced as the incorrect response. Investigating this consistent error further, we discovered that while there was no reduction or increase in the absolute number of product type errors over the learning trials, there was a significant linear increase across trials in the proportion of LCM error trials that showed this error, F=14.93, p=.002 (indicating the misconception was maintained despite an improvement in other error types). To understand the size of this effect, note that LCM errors were 23.5 % product misconceptions during the first 8 trials on average, increasing to 33.5 % product

misconceptions during the second 8 trials. Some students produce the product error with high frequency. In fact, 30 of the 255 students produced the error 3 or more times despite the fact that they got immediate feedback of the correct answers after they made these errors.

These results imply that Product items should be avoided (as we discuss in depth in the discussion section), at least until the product misconception is no longer produced for the LCM items. One possible reason for the failure to remediate the misconception shown in our data is the tendency for Product items to reinforce the incorrect strategy for true LCM problems, thereby failing to prepare students for solving these "true" "LCM items. Of course, we could suggest that this misconception is due to students' poor metacognitive analysis of their errors, but this explanation may also lead to us to conclude that reducing the number of product problems would provide a clearer message about errors, thus scaffolding these metacognitively underperforming students more effectively. Similar misconceptions in student learning about negative signs and equals signs have been shown to cause decrements in both performance and learning when they are not specifically addressed by the instruction (Booth and Koedinger 2008).

We can also see that the model reveals a much lower LCM-to-LCM item-type learning rate for failures as compared to the learning rate for successes (.420 vs. .075 as shown in Table 5). The success learning parameter is more than five times the failure learning parameter and this implies that students are only learning well if they already have some initial proficiency. The low rate of learning from failures further indicates that the feedback/review is only causing weak changes in future performance. As we will discuss in the conclusion, such results suggest that these item-types need better feedback support. It seems that merely providing the correct responses (the method used in this instruction) is insufficient to allow students to infer how to solve the LCM items. Note how the CPFA analysis method makes the need for better feedback abundantly clear, and it differentiate cases where feedback is useful, e.g., Product failure learning does much better relatively speaking than LCM failure learning, presumably because the answer allows much clearer inference of the solution procedure when that procedure is merely multiplication.

Because of the pronounced transfer effect for practices 1 to 2 compared to overall, we also computed the model restricted to data from trials 3 to 16. This model still had significant success transfer from LCM to Product at .106, p=.019. More interesting, however, was that all of the failure learning and transfer parameters were not statistically different from 0 at p<.05. For example, failure transfer from LCM to Product was .006, p=.88. Interestingly, failure transfer from Product to LCM was almost significantly negative -.11, p=.057. It seems plausible that this result helps explain the increased proportion of misconceptions produced (despite the strongly significant learning) because in this case students using a shallow strategy could easily infer (incorrectly) that all one needed to do was multiply to solve problems in the set. Generally, the restricted model results suggest that the failure feedback we gave was not cumulative and suggests that at the very least, variability of feedback may be necessary if a student keeps failing. The data implies that repeating the same feedback type does not provide sustained assistance for the student. This result shows how CPFA can be used flexibly by applying the model across subsets of data to see how transfer changes across time.

Word vs. Story Item-Types Data and Model In the previous example, the *harder* LCM item-type fostered a more general kind of learning. In contrast, in the case of Word vs. Story item-type comparison, the *easier* Word item-type appeared to better drive transfer (Fig. 2). While transfer given Word to Story item-types (dashed line between Story and Word-Story) was significant, the Story item-types did not show significant transfer to Word item-types. A summary of parameters for the CPFA model fit on the Word and the Story item-types is given in Table 6.

Word vs. Story Item-Types Discussion Transfer given Story item-type to Word item-type is non-distinguishable from zero for both successes and failures. From Word item-types to Story item-types, however, there are significant effects. Failure-driven transfer given Word to Story items is almost twice as strong as success-driven transfer (.176 vs. .083 respectively). While this difference was not significant, it may imply that in this case, perhaps due to the simplicity of review for Word items, students were able to map the analogous procedures to Story items more easily than when those same procedures were simply practiced and not reviewed. As in the case of Product and LCM item-types, transfer was asymmetric. We collapsed the success-failure distinction using a four-parameter model (for A→A, B→B, A→B and B→A trials) rather than an eight-parameter model. In this aggregate model (Story→Story=.166***, Word→Word=.303***, Story→Word=−.001 and Word→Story=.115***), we see that the pattern is very similar to the averaged parameters without success/failure categories, which helps us to confirm the addition of the success/failure parameters is not causing over-fitting.

Again we computed the model restricted to data from trials 3 to 16 because again we saw that transfer was much reduced for the overall results in Fig. 2b. Again we saw that significant transfer was maintained for success for the Word items, with a similar magnitude, .084, p=.023, while failure transfer was not significant any longer, p=.468. These results strengthen the argument that students need different/non redundant forms of feedback for later failures if we wish to have continued learning from

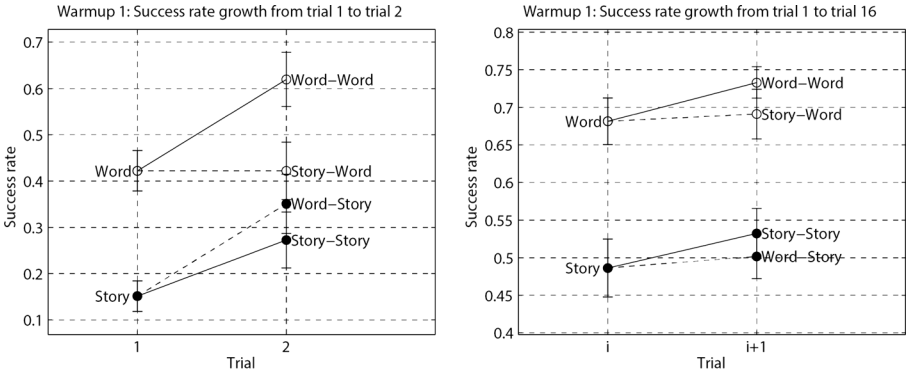

Fig. 2 Success rate growth for "Word" and "Story" items in warm-up 1: **a** from trial 1 to trial 2, **b** from trial 1 to trial 16 (averaged). 1 *SE* error bars

Table 6 Summary of the Word vs. Story item-type CPFA model parameters for warm-up 1 (including the full dataset shown in Fig. 2b)

Param.	val002E	p-value†	Param.	val.	p-value†	Note
β_0	−.983	.000***				Overall intercept (Story item-type)
β_A	1.118	.001**				Intercept modifier for Word item-type contrast level
γ_{AA}	.387	.000***	ρ_{AA}	.160	.005**	Learning from Word to Word
γ_{BA}	−.029	.556	ρ_{BA}	−.011	.811	Transfer from Story to Word
γ_{BB}	.335	.000***	ρ_{BB}	.017	.699	Learning from Story to Story
γ_{AB}	.083	.026*	ρ_{AB}	.176	.001***	Transfer from Word to Story
σ^2_u	.549					Variance of student random effect intercept
σ^2_e	.569					Variance of item random effect intercept

† Significance codes: . -$p \leq .1$, *-$p \leq .05$, **-$p \leq .01$, ***-$p \leq .001$

Success (γ) and failure (ρ)

failures. Interestingly, we do not see a similar pattern for success, which has more sustained benefits just as we saw for the Product/LCM contrast.

The greater difficulty of the Story item-type than the Word item-type for this least-common multiples context is consistent with a past finding (Koedinger 2002), but inconsistent with other contexts (Koedinger 2002; Koedinger and Nathan 2004). Given the greater difficulty of Story items in this context, we might speculate that this item-type causes excessive cognitive load that prevents students from inferring how to arrive at the correct answer upon review (Sweller et al. 1990). This argument is supported support since we also see no learning following failures with the story items. A similar explanation might be that the simplicity of the word problems allows an easier inference of a general rule for the item-type (Son et al. 2008). This argument is supported by the fact that we do see good success learning of Story items, which implies it is not a problem with learning per se, but rather a problem with the process of inference after failure.

This analysis has implications for improving instruction with these items, since the data show that the Story items are quite poor in their present state, since even when they are answered correctly, they do not transfer. To the extent that noticing the opportunity for transfer given a prior example is about noticing the analogy (cf., Singley and Anderson 1989), we might suppose that a hint given during feedback that the LCM procedure is applicable in the Story items might cause people to engage with the Story item feedback in a way that was later transferable (e.g., by applying the LCM schema). Using a hint in this fashion is similar to Gick and Holyoak's (1983) work showing how hinting about the relevance of the source analogy improves transfer performance, but in this case the hint might foster "analogical abstraction" in the encoding, that, at least hypothetically will provide benefits to improve future by allowing more general features to be added to the representation learned (Gentner et al. 2009).

Warm-Up 6 Example

Warm-up 6 addressed fraction addition items. Each item presented two fractions and students were asked to add the two fractions and, without simplifying, to produce either

the numerator or the denominator of the result. We defined two item-type splits. The first split, SameDen/DiffDen, distinguished whether the two fractions presented for addition had either the same or different denominators. The second split, AskNum/AskDen, is not analysed here because there was only non-significant symmetric negative transfer between the items that asked the student to produce the numerator of the sum and items that asked students to produce the denominator of the sum. Table 7 has examples for each item-type category and the number of item instances for it. Before building the CPFA model we filtered an original warm-up 1 dataset of 3208 data points from 225 students down to 3065 data points from 213 students by removing students that demonstrated 0 % performance (6 students) and 100 % performance (6 students).

Figure 3a is a depiction of the success rates' changes from trial 1 to trial 2 in the data from warm-up 6 for SameDen and DiffDen item-types. DiffDen (different denominator) items are harder, starting with a .14 success rate on the first trial. SameDen (same denominator) items are easier, starting with a .41 success rate. From trial 1 to 2 we see that SameDen is benefitting DiffDen, more than the reverse, and when all of the trials are averaged up to trial 16 (see Fig. 3b), the pattern is similar, but this suggestion of early transfer from SameDen-DiffDen is much diminished. Table 8 provides a summary of the parameters for CPFA model of SameDen and DiffDen item-types.

SameDen vs. DiffDen Item-Types Discussion Both SameDen and DiffDen items show strong success-learning coefficients of .407 and .420 respectively (see Table 8). Learning from failure in both cases is quite modest at best (coefficients .071 and .062), and not significant. The transfer situation is interesting. Easier SameDen items do not transfer to harder DiffDen items for either prior successes or failures (however, note the anomalous early transfer in Fig. 3a. In contrast, prior practices of DiffDen items have opposite effects on SameDen items, depending on success. Success with DiffDen items results in significant transfer to SameDen (.146 coefficient), while failure transfer from DiffDen is significantly negative with similar magnitudes (−.140). The result for success seems somewhat sensible, at least in retrospect, because, similar to the LCM-Product contrast in warm-up 1, DiffDen item solution passes through a context where

Table 7 Item examples and corresponding factors for warm-up 6

Example item texts	Factor 1	Factor 2	No. of items
What is the denominator in the solution to 4/5+2/5? (do not reduce or convert to a mixed number before answering)	SameDen	AskDen	6
What is the denominator in the solution to 1/5+2/3? (do not reduce or convert to a mixed number before answering)	DiffDen	AskDen	6
What is the numerator in the solution to 4/5+2/5? (do not reduce or convert to a mixed number before answering)	SameDen	AskNum	6
What is the numerator in the solution to 1/5+2/3? (do not reduce or convert to a mixed number before answering)	DiffDen	AskNum	6

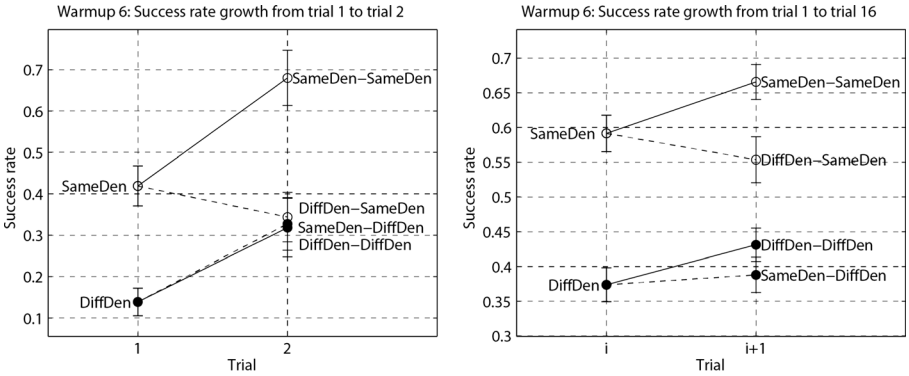

Fig. 3 Success rate growth for "SameDen" and "DiffDen" items in warm-up 1: **a** from trial 1 to trial 2, **b** from trial 1 to trial 16 (averaged). 1 *SE* error bars

the problem is set up as a SameDen item. However, failure with DiffDen does not transfer to SameDen like LCM failure transferred to Product items.

This negative failure transfer for DiffDen items may be due to the extra steps (in the case of a numerator response that requires conversion of numerators and addition) or due to the increased amount of extraneous information (in the case of denominator responses where the numerators were given but not needed to solve the problem) that occurs during feedback. Both of these problems are overcome (by definition) if the response is a success (explaining positive success transfer), but both of these problems increase the cognitive load when trying to learn from the failure (Sweller, et al. 1990). In contrast to DiffDen items, SameDen items do not appear to cause any transfer at all, perhaps because SameDen items don't practice understanding within a DiffDen context, and so this learning isn't cued by DiffDen items. Unlike for transfer from DiffDen to SameDen, the model showed

Table 8 Summary of the SameDen vs. DiffDen item-type CPFA model parameters for warm-up 6 (including the full dataset shown in Fig. 3b)

Param.	val.	*p*-value†	Param.	val.	*p*-value†	Note
β_0	−1.304	.000***				Overall intercept (DiffDen item-type)
β_A	1.174	.004**				Intercept modifier for SameDen item-type contrast level
γ_{AA}	.407	.000***	ρ_{AA}	.071	.166	Learning from SameDen to SameDen
γ_{BA}	.146	.013*	ρ_{BA}	−.140	.002**	Transfer from DiffDen to SameDen
γ_{BB}	.420	.000***	ρ_{BB}	.062	.189	Learning from DiffDen to DiffDen
γ_{AB}	−.020	.637	ρ_{AB}	.003	.962	Transfer from SameDen to DiffDen
σ^2_u	.569					Variance of student random effect intercept
σ^2_e	.814					Variance of item random effect intercept

† Significance codes: . -$p \leq .1$, *-$p \leq .05$, **-$p \leq .01$, ***-$p \leq .001$

Success (γ) and failure (ρ)

that the SameDen practice was not applicable to DiffDen (with the exception of the effect on early practice as shown in Fig. 3a).

Again we looked at the model restricted to trials 3 to 16, in part hoping to provide some insight on why the data showed early transfer benefit of SameDen to DiffDen, despite the model result. Again we saw that the restricted model showed no learning for failure. For transfer we now found a -.11 failure transfer coefficient from SameDen to Diff Den (p=.028) while DiffDen to SameDen was about the same as before for failure at -.171, p=.0002. Again this result suggests failure learning was not sustained for our feedback. This result also implies that first trial transfer benefits for SameDen items did not get detected by the model because of the later negative failure transfer effects of repetition of SameDen items. One explanation for the negative transfer in later practice comes from our analysis of misconceived responses. Similar to the case of LCM errors where students gave a Product response, there was also strong indication that some students may not have been differentiating the problems, since 33 % of numerator errors for DiffDen items were due to students producing the simple sum of the numerators, which would be correct if the denominators were the same. This percentage did not change significantly as a linear function of trials. This complex pattern may imply that some students are scaffolded by the SameDen problems, but that this occurs quickly in the first couple trials, while other students may not recognize the differences between the problems and so incorrectly transfer what they learn after failing SameDen problems to their work DiffDen problems.

To check the model stability we collapsed the success-failure distinction using a four-parameter model. In this aggregate model (SameDen→SameDen=.240***, DiffDen→DiffDen=.228***, SameDen→DiffDen=−.023 and DiffDen→SameDen=−.024), we see that the pattern is very similar to the averaged parameters without success/failure categories, which helps us to confirm the addition of the success/failure parameters is not causing over-fitting. This result helps illustrate how the performance tracking in CPFA provides an advantage. In this 4-parameter model we see no transfer effects at all while the CPFA model provided a much richer story suggesting that the problem is not with lack of transfer for DiffDen problems, but rather a problem with feedback in particular. This is a very different conclusion with important implications for improving the instructional design, which the model discovery process with CPFA has revealed.

Discussion

Summary

Using data we collected in a randomized classroom experiment, we showed how the contextual performance factors analysis (CPFA) model can be used to analyse how effective individual tasks are at promoting learning or transfer. We described differences between pairs of item-types, but given enough data to accurately identify the parameters, the method could be applied to situations where more than two item-types are practiced in a continuous random sequence. These sorts of multi-item analyses can be conducted by testing transfer between different A-B halves of the entire data set (as we demonstrated for Warm-up 1), or it could be used to look at the pairwise interactions

for multiple item-types crossed pairwise (e.g., A-B, B-C, A-C). The CPFA method assumes only that prior instructional events the student participates in can be categorized individually and the model uses these categories to determine the quantitative effects of these prior events. In addition to categorizing based on the context of practice (the item-type), the model categorizes whether prior events were correct or incorrect. This logic of counting categories of prior events is well established (Beck and Mostow 2008; Scheiblechner 1972; Spada 1977), and recent work has looked at different instructional manipulations (e.g., hints in problems vs. no hints) as categories of prior practice in a similar way (Chi et al. 2011) as the current paper considers contextual factors.

The CPFA model developed out of the data analysis, but we are certainly not the first to model asymmetric and negative transfer effects. For example, in the ACT-R (Adaptive Character of Thought - Rational) (Anderson and Lebiere 1998) theory (Anderson and Lebiere 1998), productions are capable of modelling such component transfer effects because they distinguish between different contexts of application for each production rule. A good elaboration of how production rules can capture such effects is in Koedinger and MacLaren (2002) where they discuss transferability of productions. In the case of negative transfer performance, they propose these behaviours are acquired when a production rule is created that only produces correct behaviour in specific contexts despite being cued in other contexts (i.e., the if-part or condition of the production is overly general).

Overly general production rules are therefore sensibly acquired, but produce problems in future more specific contexts. For example, in the case of students producing the product when the LCM was less than the product, we can suppose there are two alternative production rules, but that for some students, the Product production rule has an overly general context of application so that its fires even when the answer will be wrong if the result is merely the product. Therefore, one task of instructional item ordering is to determine when such overly general rules might be supported by a task and fix, remove or reorder such tasks. In warm-up 6 we saw another possible example of this issue. In this case, failure on the DiffDen problems might be causing learning that is overly generalized from the DiffDen case to cause greater failure for SameDen problems (i.e., people trying to convert numerators or denominators when the denominator was already common). This over-generalization might be caused, in part, by the minimal (correct answer only) feedback that was only given to the student for 18 s. While DiffDen review caused borderline (non-significant) learning for DiffDen problems, the time and load limitations of this review may have resulted in the negative effect on transfer to SameDen items observed (perhaps by causing students to think the SameDen items required more complex procedures also).

In Depth: Using the Model to Optimize Learning Objects

The CPFA model provides information about learning and transfer that can be used to make instructional design decisions by guiding item creation and selection according to what transfers well. The CPFA mathematical model not only provides a way to test the adequacy of item-types designed for transfer, but can also specifically pinpoint in what cases (e.g., success or failure) the learning object items are lacking or providing efficacy for learning or transfer. This specific understanding of what needs to be fixed

may be quite helpful to differentiate, for example, when to build in better review after failure or when to add self-explanation prompts to have students explain successes. What we are arguing is an approach akin to discrimination analysis in item-response theory. Discrimination analysis is the process of examining the discrimination parameters found for individual items in a set of items for a test. In such a case, some items will essentially split the students into groups (positive discrimination is when the good students get the item right, and poor students do not). In such an approach, the discrimination parameter helps the test designer determine which items to remove because they only increase the length of the test and do not improve accuracy of the identification of student ability. Similarly, analysing the item difficulty parameters allows the removal of items that are too hard and too easy and allows the designer to ensure a good spread of items across the range of possible difficulty. Just so, CPFA allows us to diagnose each item on a variety of learning-related indices that we might think are relevant.

For example, the results of the CPFA analysis of warm-up 1 show several categories of practice that lead to poor subsequent performance. We can see which and how many of the various categories of practice need redesign according to the model. To begin analysing this it makes sense to group categories, for example we note that Product to LCM transfer is always weak (i.e., for both success and failure), and we can note that Story to Word transfer is always weak (i.e., for both success and failure). Further, we can note single category issues with learning, so we see that LCM learning from failure is relatively low (for a learning parameter, which we might expect to be larger on average than a transfer parameter) and Story learning from failure is near 0.

This analysis translates to recommendations for item improvement. Beginning with the product items we must ask the controversial but warranted question, should they be used at all? From the perspective of the model, and the high number of multiplicative responses for items where the solution was less than the product, it seems these items should not be shown to students early in the process of learning this skill. It seems that these simple items, and the fact that they are easy to answer, biases people strongly to think that the task of computing the LCM is merely multiplication. Using these product items therefore may entrench the student in a misconception about the nature of the general LCM operation. In contrast to Product shortcut problems, LCM problems may begin with students computing the product, before they check to see if multiples less than the product exist. Practicing the first step creates transfer, while the second step does not cause overgeneralization, so transfer occurs. There seems no reason to use the product items.

Other steps less radical than removing this basic item type might be taken. One option would be to explicitly make the distinction between product and LCM types clear to students by making the two steps explicit, so for both problem types, the student would be asked 2 questions in succession, first, “what is the product?” and then next for the item, “what is the LCM?”. This problem structure would not remove the product problems, but would make quickly explicit the difference with LCM, since students would be able to compare the results of the two problem types explicitly, within a trial, rather than do the much more difficult cross trial inference about the differences between the unlabelled trials.

A problem this change would not fix is the poor results for LCM item failures. While the CPFA model shows how success quickly improves performance, failure has little

effect, and so we must suppose the feedback is weak (the feedback only showed the correct answer). In a case like this, the problem seems merely that the answer is left unexplained for the student. So we might suppose we could improve results here by simply making the feedback more informative, e.g., by saying, in the case of 4 and 6 for instance, that 12 is the least common multiple because it is the LEAST number that you can divide both 4 and 6 into without a remainder. Such feedback, which both explains the reasoning and shows the procedure, should allow motivated students to succeed in this category of practice.

Next we need to reflect on the Word-Story contrast differences, where the CPFA model indicates the Story items are learned poorly, since in 3 of the 4 story practice effect categories there is null benefit. Only when people understand the task and can complete it do they improve, but that improvement seems very constrained to the Story category of items. In this case, we argue that both improving success after failure and improving transfer probably involve giving the student more information about the problem structure so that upon failure they have an opportunity to see how they should have broken the problem down. So, if they respond incorrectly, the feedback may be. "No the answer is not X, for this kind of problem, think of the list of days each girl visits her grandfather as a list of multiples, e.g., 2, 4, 6. Numbers that are shared in the lists for each girl are common multiples, the least number in both lists is the least common multiple." This sort of detailed feedback seems like it should improve learning and transfer in failure.

This still leaves the problem of near 0 transfer for success with the Story items that the CPFA model detected. Here the issue may be that students who succeed on these situated Story problems may often make up natural methods like counting, without connecting those methods with the general problem type. It seems that if people saw this connection the success learning would be more likely to transfer. With this in mind we would need to add some intervention to make sure that the success is properly self-explained, perhaps with a secondary question following the success, which would ask for the same 2 numbers, "what is the LCM?" This secondary item seems trivial, but by connecting up the ad hoc procedures the students may have applied to the Story items with the terminology in the Word items it may be possible to help the students learn the abstract problem structure contained within each successfully solved contextualized problem.

Moving to warm-up 6, while it seems clear that overall this CPFA model of the data revealed a need for more rich feedback for failures, perhaps using worked examples (Atkinson et al. 2000), there was a particular problem with DiffDen items, which resulted in negative transfer to SameDen items. In contrast, we might hope for transfer in this case to be positive if students fully understood that the SameDen problems were subsumed by the procedures needed for a DiffDen item. In other words, since a DiffDen item is solved by converting to a SameDen problem first, DiffDen problem success should cause transfer to SameDen. However, if students do not learn that SameDen problems are just a simpler subgoal of DiffDen problems, they may over-generalize a more complex strategy to find a common denominator strategy (e.g., by multiplying the denominators) from DiffDen to SameDen problems. This over-generalized strategy will increase the error date on SameDen problems, yielding negative transfer. Learning may be improved with more explicit instruction on this relationship (Gick and Holyoak 1980). For example, failure feedback for DiffDen items

could explicitly use a worked example to highlight the step that arrives at the SameDen subgoal (Atkinson et al. 2009).

In general, we have shown CPFA's ability to detect the asymmetry and the type of learning interaction (success or failure) that drives the transfer is an important property, powerful for diagnosis and shaping of math curriculum item design, especially when the transfer or learning detected is weak, negative or asymmetric. Table 9 presents an overview of the asymmetric transfer effects discovered by the CPFA model. These values were computed by absolute value of the difference between the transfer parameters for each warm-up. The table shows that the asymmetry of warm-up 1 (LCM/Product) was not individually significant for success or failure, despite the fact that when the model was aggregated, the main effect of success and failure summed was significantly asymmetric (t=2.3, df=247, p<.05), with transfer from the LCM item to the Product items. Results for the warm-up 1 Story/Word factors were similar, but the asymmetry was strong enough to be at least marginally significant for success and significant for failure, with transfer best for the Word items. Results for warm-up 6 showed asymmetry both ways also, but for success, this transfer favoured the DiffDen item, whereas for failure this transfer favoured the SameDen items (which did not cause a negative effect like the DiffDen failure.)

These results are tempered by a general trend that was observed by fitting the model to subsets of the data from trials 3 to 16. In these cases we found consistent evidence that the failure transfer when it was positive was not sustained. While the significance of success for learning was maintained in these tests, answer review following failure to produce the correct answer only helped early on. In other words, multiple redundant successes appear to be consistently associated with learning while similarly redundant failures provide no help. This interesting result seems to strongly support the arguments of both traditional advocates of drill who have proposed that such practice needs to be error free (Skinner 1958), and at the same time, the result suggests that early in learning when failure is frequent, it may be essential to provide rich and varied instruction (i.e., Ainsworth 1999) since our results show doing the same thing for each failure skill repetition is clearly ineffective. This means that the results support the idea that deep conceptual learning is a prerequisite for successful practice, since unless a skill is innate, our results indicate it must pass through a stage of gradual learning with frequent failures that require varied feedback to promote learning.

Despite the weak results for failure learning we can still illustrate the models utility in making specific predictions once it has been tuned for a specific task. Let us consider the case of our recommendation that simpler practice with Product LCM items should

Table 9 Summary of asymmetric transfer effects discovered in warm-ups 1 and 6

Warm-up	Symmetry difference for success transfer (CPFA) †	Symmetry difference for failure transfer (CPFA) †
1 LCM/Product	.074 (t=1.178, p=.239)	.096 (t=1.292, p=.196)
1 Word/Story	.112 (t=1.690, p=.091*)	.187 (t=2.513, p=.012**)
6 DiffDen/SameDen	.165 (t=2.182, p=.029**)	.143 (t=1.921, p=.055*)

† Absolute value of item-type A to item-type B transfer parameter minus magnitude for item-type B to item-type A transfer. (** p<.05, * p<.10)

be skipped or held for later due to the problems with transfer and misconceptions. In this example, given LCM starts at a logit of -.919 the typical student will need about logit (3–.919=3.919/.42) 9.3 successful practices with LCM items for them to reach about 95 % correct (a logit of 3). At this level the lack of frequent errors should make it optimal to then introduce Product items, which would then start at logit 1.2 (−.919+ 1.25+9.3*.094=1.2) and need to gain 1.8 (3–1.2) to reach the 95 % correct level for product, which requires 8.65 (1.8/.208) more successful practices. In contrast, if we practice product for (3-(−.919+1.25)=2.669/.208) 12.8 more practices to arrive at 95 % correct, we must still follow it by the same 9.3 successful LCM practices. Thus, we save roughly 4 Product practices if we follow the recommendation to first practice the harder LCM items. In this example successes were assumed to simplify the math, but similar analysis can be done for combinations of success and failures using Monte Carlo simulation. In such a case there would be a very similar advantage, since the model predicts asymmetric learning from failure as well. However, again, as the preceding paragraph suggested, failure learning past trial 2 may provide little benefit for our specific task given the shallow feedback.

Qualitative Comparisons with Q-matrix Method

Our quantitative model comparisons (not shown here, but see Pavlik Jr.et al. 2009) suggest that the Q-matrix PFA skill model discussed earlier and the CPFA model are quite similar in overall prediction accuracy given the data we have analysed. However, this is not to say the models are similar qualitatively. To show these qualitative differences we simulated from each of the models given the fixed effect parameters estimated from the data (not including the random effect component). Despite the lack of difference in quantitative fit with our data, this comparison is useful, since other researchers (e.g., Bassok and Holyoak 1989) have established that in certain cases it can be crucially important to have a model that captures asymmetry of transfer. An example of these qualitative differences are shown in Fig. 4 below, which compares the QPFA (dashed) and CPFA (solid) models for the warm-up 1 case with Story items or Word items. The QPFA refers to the standard PFA model, which uses a Q-matrix to specify the assignment of needed skills to item types (Pavlik Jr.et al. 2009). Unlike CPFA, this QPFA model captures transfer to the extent that the two item types share a needed skill. The key point to note is that QPFA lines in each figure (dashed) are equivalent for the two directions and appear to merely average the different results from the CPFA model. In other words, Fig. 4 shows how QPFA is unable to capture asymmetric transfer in this situation. That CPFA model better models asymmetric transfer can be seen in the contrast between the upper graph of Fig. 4, where Word to Story transfer is upward sloping in the solid CPFA lines (positive transfer), and the lower graph, where Story to Word transfer is flat (or trending downwards) in the solid CPFA lines. QPFA is equivalent, appearing to merely average over the asymmetry.

Another example from warm-up 6, Fig. 5, shows an interesting situation where, according to the CPFA model, we see almost no effect of the same denominator items on solving of the different denominator items, but in the reverse we see a strong effect of different denominator items on same denominator items. Success on the harder different denominator items leads to practice that benefits the same denominator items. On the other hand, failed different denominator items actually predicted a reduction in

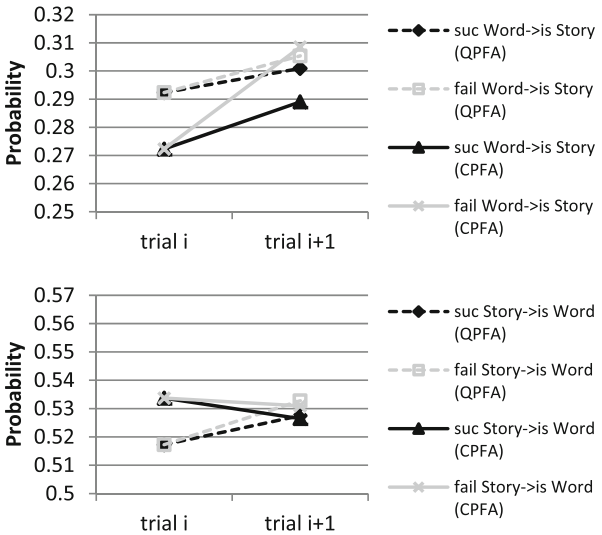

Fig. 4 Fixed effect model for trials 1 and 2 of warm-up 1 Story/Word comparison (uses Table 4 parameters). *Dashed lines* use Q-matrix PFA, and *solid lines* show CPFA

performance for same denominator items (p=.002). Failure learning required a challenging inference of how to compute the least common denominator and additionally, how to compute the summed numerators in half the items (AskNum). Students' inductive learning from examples may cue into surface features (e.g., "all the numbers [numerators and denominators] change") in a way that produces errors (negative transfer) on same denominator items. Only CPFA detected such a negative effect in this case.

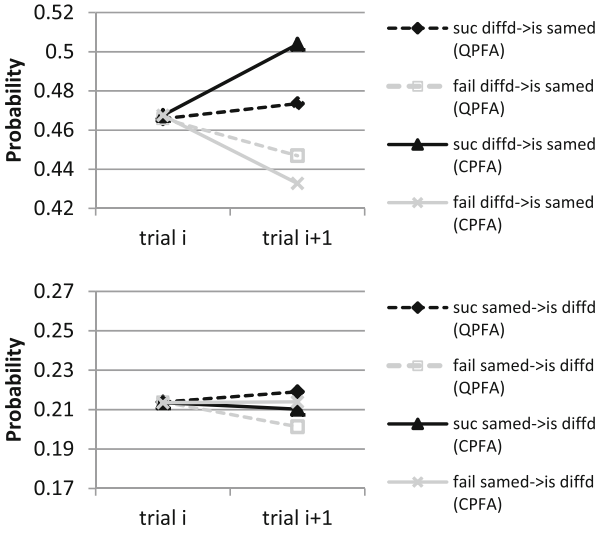

Fig. 5 Fixed effect model for trials 1 and 2 of warm-up 1 Same Denominator/ Different Denominator comparison (uses Table 8 parameters). *Dashed lines* use Q-matrix PFA, and *solid lines* show CPFA

Conclusions

CPFA (Contextual Performance Factors Analysis) is a data-driven, theory-independent method that provides a rich starting point in the process of learning theory verification and in the development of better learning objects. We demonstrated that the CPFA method of analysis gave us a way of understanding the data that leads directly to design decisions. By providing 8 event types to analyse and contrast we can quickly see five main features of the learning:

1) *The amount of learning*. CPFA measures the amount of performance change as a function of practicing the same item. The overall learning is clearly shown by how positive the learning rate parameters (for both success and failure) are for an item.
2) 2. *The amount of transfer*. The transfer measures are perhaps the most important, since if these values are low, the item type is seen as a special case and any of the learning that occurs may be bound to the context it was learned in. If these values are negative, it implies the item type is likely related, but related in a way that fosters a misconception in the other item-type.
3) *The effect of successful practice*. In addition to seeing the results of practice in learning and transfer contexts, we can also evaluate the effect of successful practice specifically by looking at learning and transfer parameters found for successful interactions. If these values are low, it implies that performing the task by itself doesn't cause learning.
4) *The effectiveness of feedback*. When students get items wrong they see feedback in most learning paradigms, and this is tracked by the failure learning parameters in our model. Good items will need to result in improved performance even if the student cannot perform the task. Further, passive review of this sort is not thought to lead to deep learning. Because of this general weakness of feedback relative to practice. These measures of feedback effectiveness are key to interpreting how the task is affecting the neediest students, since they produce more errors and get more feedback.
5) *The amount of transfer relative to learning* (e.g., $\gamma BA/\gamma AA$). Considering these ratio is useful in indicating how general the learning is. If these value are near 1, it indicates transfer is as strong as learning. Certainly we would like to see this value be high, but when the value is very high it may also be indicating that the difference between the two item types is very minimal. Indeed, we might suppose that for any target item-type we desire to transfer to that there is some optimal amount of transfer relative to learning. We are looking for situations where transfer is a large fraction of learning, but also where such transfer occurs between items that are not nearly identical. In other words, design of systems should balance transferability with novelty when new item-types or activities are introduced. Items need to be similarly related enough to result in transfer, but different enough that the new skill provides novel benefits.

While we have demonstrated CPFA in the context of single step problems, the method could also be used for multistep problems by creating two or versions of multistep problems where there was some question about the optimal ordering of two steps. Thus, in a problem with steps A, B, C, and D, and one might test step orders A-

B-C-D and A-C-B-A. This would give a quick and simple report for the 4 transfer parameters, showing which of these conditions was producing better learning and transfer. This analysis could reveal a number of issues, since even without the non-transfer learning conditions the model can measure asymmetric transfer and determine the effect of step success or step initial failure. Indeed, the applicability of CPFA may be very broad due to its generality. Currently under investigation is the extent to which a CPFA type model is a good representation of semantic learning as measured by cloze sentences. Using a collection of sentences about statistics we may be able to measure the extent to which related sentences potentiate learning of other sentences. So, for instance the two sentences: "a distribution describes the likelihood that observations will occur within any range of values," and "the normal distribution has 68 % of its observations in the range between −1 and 1 standard deviation," are clearly related and we might suppose that there is some transfer from reading one to support reading/encoding of the other, which CPFA could measure. However, this reveals a limitation of the method, since if there are more than 2 related items, even if they are not part of the model, the other related items may add considerable noise to the effort to tease apart a specific relationship.

CPFA is an analytic tool that supports instructional material improvement as well as the development of process-oriented transfer theories by empirically measuring different kinds of transfer between equivalence classes of stimuli. Unlike psychological or cognitive science theories of transfer, it makes no attempt to explain the processes and structures underlying practice. CPFA helps in interpreting empirical results that can be used to test theories of transfer. For example, consider Ohlsson's theory of transfer and learning (Ohlsson 2011), which proposes mechanisms of transfer that involve the specialization of general strategies. In this theory, for example in the LCM task, we may have explained the failure of Product practice to transfer as due to a lack of specialized version of the general skill of multiplication. The theory would probably claim that the Product practice transferred poorly because it involved only the general multiplication rules. In contrast, as a specialized version of Product skill, the LCM skill practice results in Product learning since Product learning is just a more general version of LCM learning. This hierarchy of strategies in the deep learning theory is similar to the way some cognitive models describe the evolution of verbal learning, as proceeding to more and more specific interpretations as the learner takes in more and more prior exemplars to form specific concepts (Gobet 1998; Thiessen and Pavlik Jr. 2013). It may be that practice tends to transfer best when it goes from more specific to more general.

Such theoretical approaches have strong utility, since they can narrow the space of possible quantitative models by describing which effects do not need to be tracked or attended to when modelling. For instance, if we could show that transfer tends to follow the patterns described by Ohlsson's deep learning theory for a domain of equivalence classes of items, then we could assume a few things. For instance, deep learning might say that the more specific LCM task is adapted from the more general product task, so this tells us that the LCM task will be learned more slowly as a function of the Product task knowledge. In other words, the LCM learning rate should be positively correlated with the Product skill logit. Despite this, because the specialized LCM knowledge is still weak, students may produce overgeneralizations where they apply the Product skill to an LCM item as we observed in the data. Perhaps the lack of transfer we observed is due to the balancing of these opposing effects. In contrast, LCM learning necessarily

engages the specialized skill, which naturally includes practice of the more general skill it was derived from.

Theories provide these richer explanations, and constrain modelling, but theories also have exceptions and require that the investigated evidence be categorized and quantified in a way that lacks theoretical presuppositions. So unless we know a theory is correct for a domain, it is risky to build it into the analytic model ahead of time. CPFA, since it is not a theory, but rather a statistical model, avoids this risk during development of models and systems by making only one very weak assumption, that is, that there exist different types of items that are related. This weak theoretical assumption means that the patterns in the model may support some theories (e.g., deep learning) or provide evidence against others (e.g., theories consistent with a simple Q-matrix model), since these theories predict different patterns of parameters.

Future work with CPFA may concentrate on improving its integration with other techniques for student learning such as IFA (Instructional Factors Analysis) to further gain details about the transfer patterns in students (Chi, et al. 2011). By combining the methods, it becomes possible to create a model that blends the mechanisms depending on the learning object from which transfer needs to be measured. This may allow us to discern important details of transfer, such as the relative effects of a hint when it appears as feedback, compared to the same hint given prior to practice as instruction. Similarly, such work will help us answer the question of whether logistic linear growth is useful to characterize the effect of feedback (or declarative conceptual learning) generally. Indeed, our results indicate that repeated feedback of the same sort behaves much differently than repeated successes. While feedback helps for failure, we saw that this effect only occurred for the initial failure trials.

Another interesting opportunity for future work involves how the methods may be used for individualizing student instruction through task/item selection depending on a CPFA model of the student. In the current paper, the problems were not adequate for such individualized instruction, since the feedback effect was not sustained. However, if we imagine that each problem had a menu of feedback statements depending on student responses and was designed to give a new perspective on or means for solution (i.e., rich hints or explanations of specific cases), it seems plausible that CPFA would be useful to track students learning and select item according to what is optimal for transfer. While it may become clear in future work how to control the parameter explosion that comes with multiple item types using the CPFA method, such scheduling would currently only be useful for units of content with only a few problem types. Future work may be able overcome this limitation if there was a principled way to triage all the possible relationships between the n skills being scheduled. Such future work might use methods such as knowledge space theory to determine which pairs of items are related (Desmarais et al. 1996; Falmagne et al. 2003), and then when a relationship is determined that relationship could be fit with CPFA parameters to understand the micro genetics of any transfer between the items. Of course this presumes that the design of the system allows for some variability in order, since if A always occurs before B, measuring B to A transfer is impossible.

Because CPFA does not require any in depth task analysis other than categorizing items, and can be implemented as standard mixed effect logistic regression, it is relatively easy to use and cost effective as a method of understanding how learning objects are being received and processed by students. A CPFA model of data in a

domain is primarily useful because of the implications of the model for designing, comparing and sequencing items or other learning objects. For this usefulness to be manifest, it is important that data input to the model fitting is unbiased in its sequence, since the statistical method assumes this unbiased item ordering. Given this unbiased item ordering the method is applicable in the very general situation where there are two or more different item-types given in a sequence. The model itself supports three possible transfer relationships between two item categories: no transfer, asymmetric transfer, or symmetric transfer. Because the model captures negative transfer, it has the potential to categorize item-types that cause or strengthen misconceptions. Because the model reveals asymmetrical and symmetrical transfer, the model can be used to diagnose whether a particular item-type teaches a more generalizable sort of knowledge or a more situated type of knowledge.

Acknowledgments This research was supported by the U.S. Department of Education (IES-NCSER) #R305B070487 and was also made possible with the assistance and funding of Carnegie Learning Inc., the Pittsburgh Science of Learning Center, DataShop team (NSF-SBE) #0354420 and Ronald Zdrojkowski.

References

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2–3), 131–152.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology Learning Memory and Cognition, 20*(6), 1322–1340.

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

Atkinson, R. C. (1972). Ingredients for a theory of instruction. *American Psychologist, 27*(10), 921–931.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214.

Atkinson, R. K., Lin, L., & Harrison, C. (2009). Comparing the efficacy of different signaling techniques. In G. Siemens & C. Fulford (Eds.), *World conference on educational multimedia, hypermedia and telecommunications 2009* (pp. 954–962). Honolulu: AACE.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Barnes, T. (2005). The Q-matrix method: mining student response data for knowledge. Paper presented at the American Association for Artificial Intelligence 2005 Educational Data Mining Workshop.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology Learning Memory and Cognition, 15*(1), 153–166.

Beck, J., & Mostow, J. (2008). How who should practice: using learning decomposition to evaluate the efficacy of different types of practice for different types of students. (pp. 353–362).

Booth, J. L., & Koedinger, K. R. (2008). Key misconceptions in algebraic problem solving. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 571–576). Austin: Cognitive Science Society.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: a simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61–100.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis - a general method for cognitive model evaluation and improvement *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 164–175). Springer Berlin / Heidelberg.

Cen, H., Koedinger, K., & Junker, B. (2008). Comparing two IRT Models for conjunctive skills. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent tutoring systems* (Vol. 5091, pp. 796–798). Springer Berlin Heidelberg.

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific-reasoning and problem-solving strategies in children. *Advances in Child Development and Behavior, 36*, 419–470.

Chi, M., Koedinger, K. R., Gordon, G., Jordan, P., & VanLehn, K. (2011). *Instructional factors analysis: A cognitive model for multiple instructional interventions.* Proceedings of the 4th International Conference on Educational Data Mining (pp. 61–70), Eindhoven, The Netherlands.

Corbett, A. T., & Anderson, J. R. (1992). Student modeling and mastery learning in a computer-based programming tutor. In C. Frasson, G. Gauthier, & G. McCalla (Eds.), *Intelligent tutoring systems: Second international conference on intelligent tutoring systems* (pp. 413–420). New York: Springer.

Desmarais, M. C., Maluf, A., & Liu, J. (1996). User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction, 5*(3–4), 283–315.

Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 103–125).

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples. *Artificial Intelligence, 41*(1), 1–63.

Falmagne, J.-C., Doignon, J.-P., Cosyn, E., & Thiery, N. (2003). The assessment of knowledge in theory and in practice. *Institute for Mathematical Behavioral Sciences, Paper 26*.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: a general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393–405.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: analogical abstraction supports relational retrieval of past events. *Cognitive Science, 33*(8), 1343–1382.

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47*(3), 196–229.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. [Journal Article]. *Cognitive Psychology, 12*(3), 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38.

Gobet, F. (1998). Expert memory: a comparison of four theories. *Cognition, 66*(2), 115–152.

Gong, Y., Beck, J., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (Vol. 6094, pp. 35–44). Springer Berlin / Heidelberg

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220–264.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Koedinger, K.R. (2002). Toward evidence for instructional design principles: Examples from cognitive tutor math 6. *Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education)* (pp. 21–49).

Koedinger, K. R., & Corbett, A. T. (2010). *The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning*. Pittsburgh: Carnegie Mellon University.

Koedinger, K. R., & MacLaren, B. A. (2002). Developing a pedagogical domain theory of early algebra problem solving *CMU-HCII Tech Report 02–100*.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: effects of representation on quantitative reasoning. *The Journal of the Learning Sciences, 13*(2), 129–164.

Koedinger, K. R., Aleven, V., Roll, I., & Baker, R. S. J. D. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education*. New York: Routledge.

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533.

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge University Press.

Pavlik, P. I., Jr. (2007). Understanding and applying the dynamics of test practice and study practice. *Instructional Science, 35*, 407–441.

Pavlik Jr., P. I., Yudelson, M., & Koedinger, K. R. (2011). Using contextual factors analysis to explain transfer of least common multiple skills. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education* (Vol. 6738, pp. 256–263). Berlin, Germany: Springer.

Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). In V. Dimitrova, R. Mizoguchi, B. D. Boulay, & A. Graesser (Eds.), *Performance factors analysis – A new alternative to knowledge tracing* (pp. 531–538). Brighton: Proceedings of the 14th International Conference on Artificial Intelligence in Education.

Peterson, L. R. (1965). Paired-associate latencies after the last error. *Psychonomic Science, 2*(6), 167–168.

Postman, L., Keppel, G., & Zacks, R. (1968). Studies of learning to learn: VII. The effects of practice on response integration. *Journal of Verbal Learning and Verbal Behavior, 7*(4), 776–784.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49–57.

Rickard, T. C. (1997). Bending the power law: a CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126*(3), 288–311.

Rickard, T. C. (1999). A CMPL alternative account of practice effects in numerosity judgment tasks. *Journal of Experimental Psychology Learning Memory and Cognition, 25*(2), 532–542.

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology Learning Memory and Cognition, 22*(5), 1281–1295.

Rittle-Johnson, B., Saylor, M., & Swygert, K. E. (2008). Learning from explaining: does it matter if mom is listening? *Journal of Experimental Child Psychology, 100*(3), 215–224.

Romero, C., & Ventura, S. (2007). Educational data mining: a survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Scheiblechner, H. (1972). Das Lernen und Losen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19*, 476–506.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects - evidence from 2nd-language word recognition. *Applied Psycholinguistics, 14*(3), 369–385.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60*(1), 607–629.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: a direct means for studying cognitive development. *American Psychologist, 46*(6), 606–620.

Singley, M. K., & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies, 22*(4), 403–423.

Singley, M. K., & Anderson, J. R. (1989). Transfer in the ACT* theory *The transfer of cognitive skill* (pp. viii, 300). Cambridge, MA, US; Harvard University Press.

Skinner, B. F. (1958). Teaching machines; from the experimental study of learning come devices which arrange optimal conditions for self instruction. *Science (New York, N.Y.), 128*(3330), 969–977.

Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review, 12*(3), 508–513.

Son, J. Y., & Goldstone, R. L. (2009). Fostering general transfer with specific simulations. *Pragmatics and Cognition, 17*, 1–42.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2008). Simplicity and generalization: short-cutting abstraction in children's object categorizations. *Cognition, 108*(3), 626–638.

Spada, H. (1977). Logistic models of learning and thought. In H. Spada & W. F. Kempf (Eds.), *Structural models of learning and thought* (pp. 227–262). Bern: Huber.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. Embretson (Ed.), *Test design: Developments in psycholgoy and psychometrics*. Orlando: Academic.

Stamper, J., & Koedinger, K. (2011). Human-machine student model discovery and improvement using DataShop. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education* (Vol. 6738, pp. 353–360). Berlin: Springer.

Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. *Proceedings of the National Academy of Sciences of the United States of America, 105*(19), 6791–6792.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. [Journal Article]. *Journal of Experimental Psychology: General, 119*(2), 176–192.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: simple mechanism to model complex skill acquisition. [Journal Article]. *Human Factors, 45*(1), 61–76.

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.

Thiessen, E. D., & Pavlik, P. I., Jr. (2013). iMinerva: a mathematical model of distributional statistical learning. *Cognitive Science, 37*(2), 310–343.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: a reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*(3), 210–221.

Thompson, C. K., Shapiro, L. P., & Roberts, M. M. (1993). Treatment of sentence production deficits in aphasia: a linguistic-specific approach to wh-interrogative training and generalization. *Aphasiology, 7*(1), 111–133.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (I). [Journal Article]. *Psychological Review, 8*(3), 247–261.

Yudelson, M., Pavlik Jr., P. I., & Koedinger, K. R. (2011). User modeling – A notoriously black art. In J. Konstan, R. Conejo, J. Marzo, & N. Oliver (Eds.), *User modeling, adaption and personalization* (Vol. 6787, pp. 317–328). Springer Berlin / Heidelberg.