CRESST REPORT 841

THE EFFECTS OF MATH VIDEO GAMES ON LEARNING: A RANDOMIZED EVALUATION STUDY WITH INNOVATIVE IMPACT ESTIMATION TECHNIQUES

AUGUST 2014

Gregory K. W. K. Chung
Kilchan Choi
Eva L. Baker
Li Cai

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

The Effects of Math Video Games on Learning: A Randomized Evaluation Study with Innovative Impact Estimation Techniques

CRESST Report 841

Gregory K. W. K. Chung, Kilchan Choi, Eva L. Baker, and Li Cai
CRESST/University of California, Los Angeles

August 2014

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C080015 to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

To cite from this report, please use the following as your APA reference: Chung, G. K. W. K., Choi, K., Baker, E. L., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

THE EFFECTS OF MATH VIDEO GAMES ON LEARNING: A RANDOMIZED EVALUATION STUDY WITH INNOVATIVE IMPACT ESTIMATION TECHNIQUES

Gregory K. W. K. Chung, Kilchan Choi, Eva L. Baker, and Li Cai
CRESST/University of California, Los Angeles

Abstract

A large-scale randomized controlled trial tested the effects of researcher-developed learning games on a transfer measure of fractions knowledge. The measure contained items similar to standardized assessments. Thirty treatment and 29 control classrooms (~1500 students, 9 districts, 26 schools) participated in the study. Students in treatment classrooms played fractions games and students in the control classrooms played solving equations games. Multilevel multidimensional item response theory modeling of the outcome measure produced scaled scores that were more sensitive to the instructional treatment than standard measurement approaches. Hierarchical linear modeling of the scaled scores showed that the treatment condition performed significantly higher on the outcome measure than the control condition. The effect (d = 0.58) was medium to large (Cohen, 1992).

Introduction

Funded by the Institute of Education Sciences, UCLA/CRESST established the Center for Advanced Technology in Schools (CATS) in fall 2008. The primary goal of CATS is to develop and evaluate the effectiveness of computer-based games aimed at improving students' knowledge of pre-algebra topics.

The allure of using computer games for learning purposes lies in the potential for games to support multiple learning outcomes while focusing, increasing, and maintaining learners' engagement in the relevant tasks. Well-designed games will be able to address key elements understood to influence learning and performance. These include (a) focusing learners' attention on the game (and thus content) for extended periods of time, (b) accommodating complex and diverse approaches to learning processes and outcomes, (c) embedding high interactivity, (d) providing appropriate feedback, (e) creating a sense of enjoyment and intense engagement ("flow" or "presence"), (f) requiring problem-solving skills, (g) providing scaffolding and adaptive challenge, (h) creating contextual learning outcomes, and (i) potentially influencing learners' self-efficacy and other affective constructs (e.g., de Freitas, 2006; Kirriemuir & McFarlane, 2003; Mayer, 2011; O'Neil, Wainess, & Baker, 2005; Squire, 2011; Tobias, Fletcher, Bediou, Wind, & Chen, 2014; Tobias, Fletcher, Dai, & Wind, 2011; Tobias, Fletcher, & Wind, 2014; Young et al., 2012).

There is growing empirical evidence that games can be effective in academic settings (Tobias, Fletcher, Bediou, et al., 2014; Tobias, Fletcher, & Wind, 2014). Educators and trainers have recognized the potential of individualized feedback and computer games for education and training since the mid-20th century (e.g., Wiener, 1954), with researchers exploring various methods to increase student engagement with subject matter using various forms of games (e.g., Donchin, 1989; Malone, 1981; Malone & Lepper, 1987; Ramsberger, Hopwood, Hargan, & Underhill, 1983; Ruben, 1999; Thomas & Macredie, 1994). A renewed interest and optimism have emerged around games for learning, particularly games that incorporate interactive multimedia afforded by rapid developments in technology (e.g., Dickey, 2005; Gee, 2003, 2004; Kafai, 2006; Kafai, Franke, Ching, & Shih, 1998; Klopfer & Squire, 2004; O'Neil & Perez, 2008).

Because of the importance of algebra and the high rates of poor performance in algebra, we focused on foundational pre-algebra concepts. Success in algebra is predicated on students developing foundational math concepts and skills. The National Mathematics Advisory Panel (NMAP, 2008) defined the Critical Foundations of Algebra as (a) fluency with whole numbers, (b) fluency with fractions, and (c) particular aspects of geometry and measurement. One of the clearest findings of the NMAP report is that students entering algebra are often underprepared. In their national sample of Algebra 1 teachers, NMAP reported that rational numbers was one of the areas that teachers reported their students being poorly prepared in.

Thus, CATS focused on instantiating foundational pre-algebra concepts in games and carried out the development and testing of those games, culminating in a large-scale multi-site cluster randomized controlled trial (RCT). The overarching research question for the RCT was: *Does playing CATS-developed rational numbers games result in more learning of rational numbers concepts compared to playing a control set of games?*

Method

Game set. Eight games were developed for this study through the process of knowledge specification, software design and testing, teacher professional development, and assessment development. A central component of the development process was establishing knowledge specifications. The purpose for developing knowledge specifications was to provide a standardized operationalization of the domain definition of the mathematical knowledge that the game, assessment, and professional development were to be designed around. Two sets of knowledge specifications were developed: rational numbers/fractions, and solving equations. The knowledge specifications served as the design framework for the

games. Four games covered fractions concepts (number line concepts, fraction addition, relationships among whole numbers and fractions using multiplication and division, direct variation) and four games covered solving equations concepts (integer operations, expressions, solving equations—conceptual, solving equations—procedural). The four fractions games were *Wiki Jones, Save Patch, Tlaloc's Book,* and *Rosie's Rates*, and the four solving equations games were *Monster Line, Expresso, Zooples in Space,* and *AlgebRock.* Table 1 shows the set of games for the treatment and control conditions. As presented in Table 1, the fractions games were implemented as treatment condition games, and the solving equations games served as control condition games. A full description of each game is given in CATS (2012).

Table 1

CATS-Developed Games by Condition

Game	Topics
Treatment condition games (fractions)	
Wiki Jones	Whole unit, numerator, denominator, and identifying fractions in the context of a number line.
Save Patch	Addition of fractions, identification of units and fractional pieces.
Tlaloc's Book	Multiplicative inverse operations involving whole units and fractional units.
Rosie's Rates	Direct variation (slope) involving relating changes in x values to changes in y values.
Control condition games (solving equations)	
Monster Line	Addition, subtraction, multiplication, and division of positive and negative integers.
Expresso	Manipulation of expressions involving positive or negative whole numbers and variables, and grouping.
Zooples in Space	Concept of equality and the use of additive inverse operations.
AlgebRock	Solving one- and two-step equations.

Design and sample. A cluster randomized trial (CRT) design was used where individuals were nested within classrooms. Sixty-two teachers were originally recruited to participate in the study. Two teachers from the same school withdrew because of technology issues and one teacher from another school withdrew because that teacher could not schedule sufficient computer lab time to complete the study. No other teachers withdrew from the study.

Twenty-four schools participated in this study, but one school was excluded from the analysis because the intervention could not be on the (obsolete) computers in that school.

Classrooms were randomly assigned within each school to the treatment and control conditions. Among the remaining 23 schools, 14 schools had both control and treatment conditions implemented, while the remaining 9 schools had either only treatment condition (6 schools) or only control condition (3 schools) classrooms. Note, however, that all the participating schools administered both the pretest and the posttest measure of fractions knowledge. The pretest and posttest measures largely overlap, that is, the vast majority of items were repeated with just a few variant items.

Classrooms were sampled from sixth grade math classes. Of the total sample, 50% of the student sample were female, 49% Hispanic/Latino/a, 24% White, 11% multiracial, 5% Black or African American, 4% Asian or Pacific Islander, 2% American Indian or Alaskan Native, and 5% of students reported "Other."

Measures. The outcome measure of fractions knowledge was developed, tested, and refined during the game testing process. Vendlinski, Delacruz, Buschang, Chung, and Baker (2010) reported that the outcome measure demonstrated high technical quality. There were 22 items on the pretest and 23 items on the posttest. The items were systematically developed from the knowledge specifications and were similar to items found in typical standardized assessments on those topics. Classical discrimination indices were adequate and the items were not overly easy or difficult for the target sample (proportion correct ranged from .4 to .7). The range of item to total score correlation was from .09 to .85. Classical reliability estimates were also moderate (.80).

Professional development (PD). Prior to using the games in the classroom, teachers received PD training on how to integrate the games into their curriculum. This training included background information on the research behind the math topics underlying the games, the common errors associated with the various concepts, the mathematical concepts covered by the games, and the linkage between the game mechanics and the mathematical operations. In teacher training sessions, teachers were randomly assigned to a condition prior to attending the sessions. Teachers assigned to the treatment condition received training on games related to fractions concepts, and teachers in the control condition received training on games related to solving equations concepts. To accommodate the availability of teachers, the three-hour PD session was split into two sessions of 1.5 hours each, if needed. Some PD meetings were conducted during school hours while others were held during after-school hours at the district office or at a school site within the district. The first part of the PD

session was designed to help teachers understand key conceptual ideas and student misconceptions around mathematics concepts in the video games. Teachers discussed general "roadblocks" to understanding and then looked at the video game to see how these math concepts were addressed in the game. The second part of the PD session focused on having teachers play the video games. We had found that many teachers do not play video games; therefore the incorporation of video games into the classroom could be intimidating or difficult to manage even for a talented math teacher. Because of the low initial comfort level, one of our goals was to have teachers play through as much of the video games as possible to both give them experience playing the video game and increase their comfort level with playing video games. The final part of the PD session focused on helping teachers link the video game to their mathematics instruction. Participants discussed their experiences playing the game, how these video games could be incorporated into instruction, and how these games might benefit students.

Procedure. The efficacy trial study required 12 instructional days (10 gameplay days and 2 testing days). Students were first administered a pretest measure (prior knowledge of mathematics, attitudes toward math, and game skill and experience). Students then played each game in a prescribed sequence for a set number of periods as shown in Appendix B. In general, students played games for at least 40 minutes per period and between two and four periods per game. After completing each game, students were administered an immediate posttest on content related to the just completed game and a game perception measure. A delayed posttest was administered a week after the last gameplay day.

Teachers also completed measures in the following sequence: Teachers provided feedback on the games during the professional development session, completed a background measure during the pretest day, kept logs of student activities and problems on each game day, and listed the topics covered between the last gameplay day and the posttest. Teachers also provided general comments on their study experience after the posttest.

Results

Methodology Research Question

The key research question in this analysis focuses on estimating the game's treatment impact on students' fractions knowledge learning outcomes. We conducted two different sets of analyses. One analysis used the classical measurement approach in which raw summed scores of pretest and posttest items were calculated, while the other analysis utilized a multilevel two-tier (MTT) item factor model in which a latent gain score was estimated. As will be described in detail in the following section, we illustrate how the MTT model

addresses four critical aspects of *conditional exchangeability* that routinely accompany analysis of multisite randomized experiments with pre- and posttests.

In the absence of conditioning on appropriate observed or latent variables in a measurement model, the observations of student performance on outcome items or tasks are correlated/dependent in four major ways, on top of the dependence of item responses themselves: (1) the dependence between the outcome constructs at each occasion due to a longitudinal design; (2) the item-level residual dependence due to repeated (pre-post) exposure to the same set of measures; (3) the practical implausibility of assuming full exchangeability of subjects across treatment and control conditions (see e.g., Lindley & Smith, 1972); and finally (4) the obvious dependence of individuals due to their nesting in sites. We argue that the traditional summed score approach (using classical test theory) or standard "off-the-shelf" Item Response Theory (IRT)-based approaches have assumptions that are inconsistent with the conditional exchangeability implied and required by multisite randomized experimental studies with repeated measures. In contrast, the MTT model embraces conditional exchangeability and specifies model features that appropriately reflect the interaction of latent variable measurement models with the experimental design. We compare impact estimates obtained from different approaches and explain why MTT modeling provides a superior solution to measurement and data analysis issues in multisite randomized trials.

Summed Score Model

The simplest and most commonly used method for scoring outcome measures is via the summed score model (e.g., Curran, Bauer, & Willoughby, 2004; Curran & Bollen, 2001; Curran et al., 2008). The ubiquity of this approach stems from the straightforward method of scoring and the long history of classical summed score based test theory in the social and behavioral sciences. As can be seen in Equation 1 below, the summed score ($\hat{Y}_j$) for person j is calculated by adding the raw item scores, where y_{ij} represents the observed item scores (e.g., 0 or 1 for binary cases; 0, 0.5, or 1 for three partial-credit scoring categories, etc.) on item i. I denotes the total number of items.

$$\hat{Y}_j = \sum_{i=1}^{I} y_{ij} \qquad (1)$$

This simple method, however, has critical disadvantages. Since this method simply adds up the item scores with equal weighting, differences in item difficulty, discrimination, and/or student guessing (among other psychometric characteristics) are completely ignored. Comparability problems arise when the numbers of observed items are different between tests or occasions, either by design or due to missing data. Yet another complicating factor is

that preassigned item weights (often by fiat, e.g., giving partial-credit scored open-ended items more weight than dichotomously scored multiple-choice items) may lead to suboptimal reliability in that the total score may in fact be less reliable than component subscores. Finally, inferences derived from the summed score distribution are sample-specific and dedicated linking/equating studies are required for generalization of the sample-based results to other populations.

Nevertheless, we utilize the summed scoring method to establish baseline results for comparison purposes. In our study, the total numbers of items for pretest and posttest are, respectively, 22 and 23. Among 22 pretest items, 17 items are dichotomously scored (0 or 1), and the remaining 5 items are partial-credit scored: two items to 0, 0.5, or 1; two items to 0, 0.3, 0.67, or 1; and one item to 0, 0.25, 0.5, 0.75, or 1. Similarly, among 23 posttest items, 17 items are scored to either 0 or 1; one item to 0, 0.5, or 1; three items to 0, 0.3, 0.67, or 1; and two items to 0, 0.25, 0.5, 0.75, or 1. Note that there are 20 common items administered in both the pretest and posttest.

Figure 1 presents descriptive statistics of raw summed pretest and posttest scores for control and treatment conditions (see also Table A1 in Appendix A). The dot and the vertical bar represent the mean and one standard deviation above and below the mean. The total numbers of students for control and treatment conditions used in our analyses are 763 and 808, respectively. In the control group, there are 709 students who have both pretest and posttest scores, 54 students missing posttest, and 36 students missing pretest. In the treatment group, there are 759 students who have both pretest and posttest scores, 49 students missing posttest, and 33 students missing pretest. The pretest mean score for the control group is about 8 and its standard deviation (*SD*) is approximately 4. Similarly, the pretest mean score for the treatment group is 8 and its *SD* is 4.1. Although random assignment was at the classroom level within each school, highly similar overall pretest means across the two conditions provide another layer of assurance that the randomization indeed led to balance on pre-treatment differences in mathematics knowledge.

The posttest mean score is higher by approximately 1.5 points for the treatment group than for the control group. The observed posttest mean scores are 10.9 for the treatment group and 9.5 for the control group. This difference is approximately 0.3 pooled standard deviation of posttest.

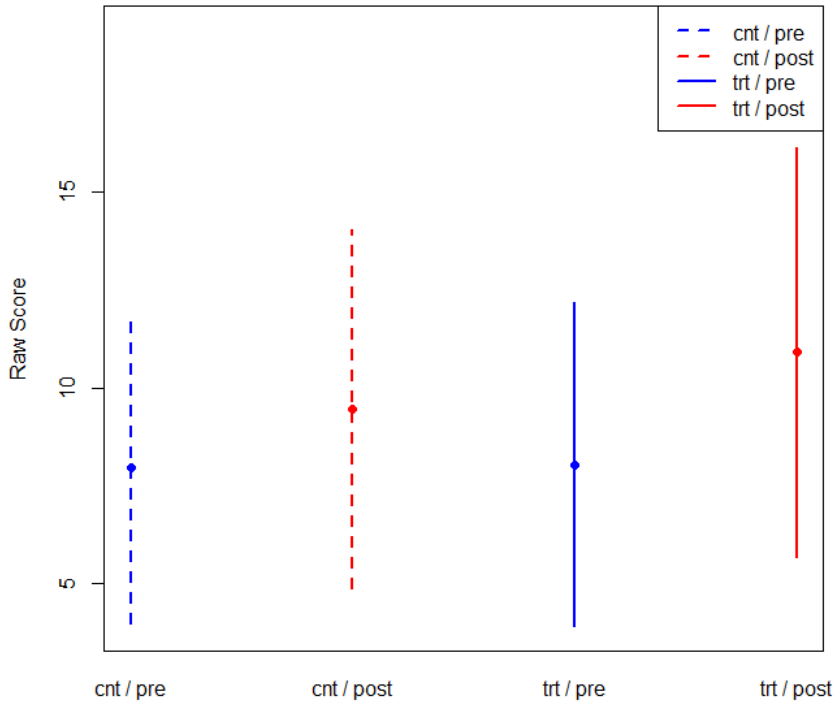

Figure 1. Descriptive statistics of raw total scores of pretest and posttest by experimental conditions.

We also examined the descriptive statistics of raw total pretest and posttest scores by schools and experimental conditions. As can be seen in Figure 2, there is some variability in pretest mean scores across schools. School 8 has a pretest mean of approximately 4 points, while School 9 has a pretest mean close to 13 points. Except for these two schools, most of the rest of the schools have similar pretest mean scores. Within-school pretest difference between the control and the treatment, which is more important in a randomized trial, is not salient. Specifically, the pretest differences between the two groups in most of the schools range within a point and half, but three schools (Schools 8, 19, and 23) show differences of approximately 2.5 points or higher.

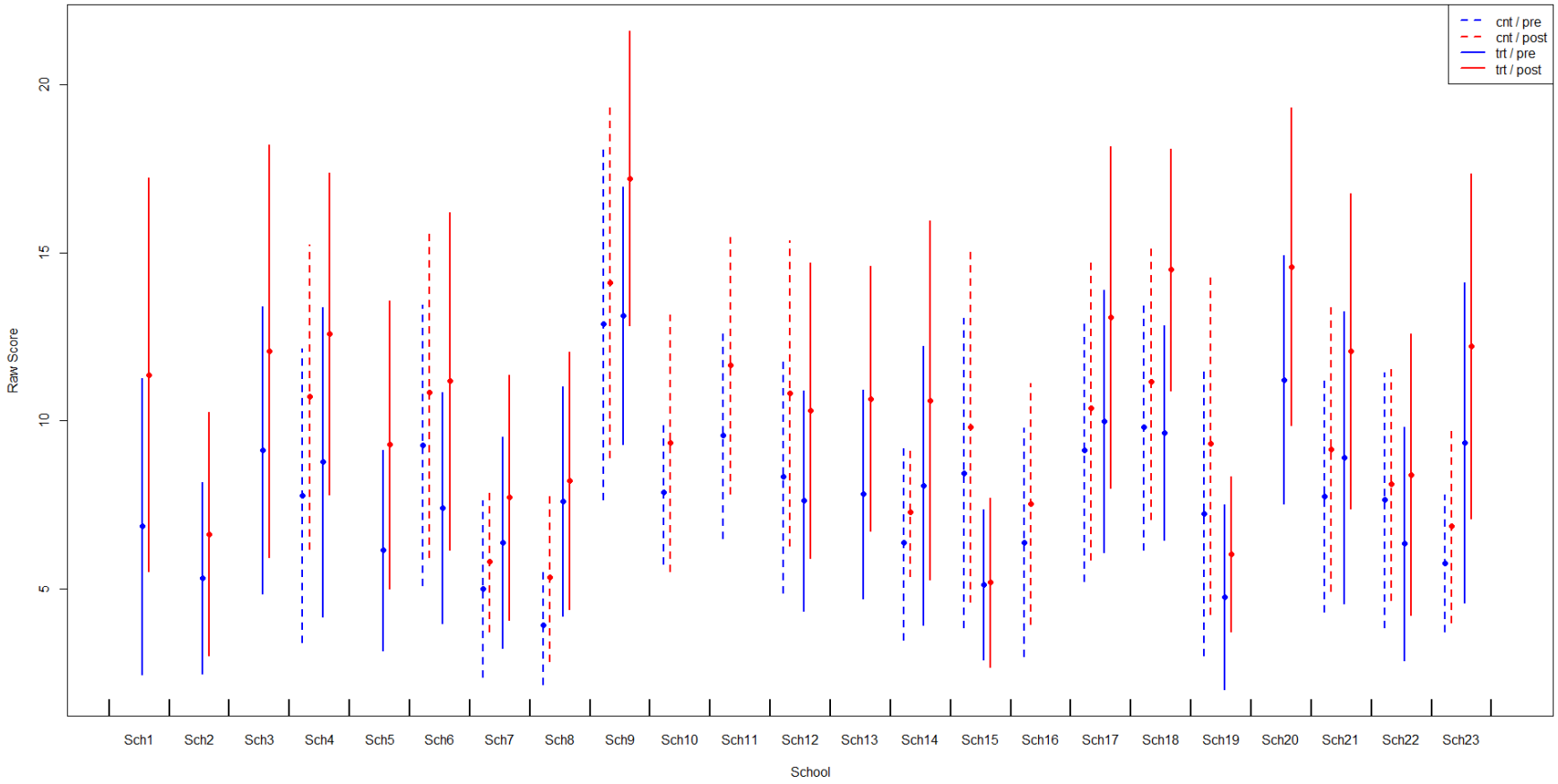

Figure 2. Descriptive statistics of raw total scores of pretest and posttest by schools and experimental conditions.

As for posttest mean scores, all the schools show higher posttest means than pretest means both in control and treatment groups. This positive increase in mean scores is partly attributable to the fact that the posttest has one more item than the pretest. However, the differences in many schools are larger than one point, and the differences are larger for the treatment groups. For example, the difference between pretest mean and posttest mean in the treatment group in Schools 1, 4, and 18 is approximately 4 points. Furthermore, the differences between posttest mean and pretest mean are larger in the treatment group than in the control group for 10 out of 13 schools which have both treatment and control groups. These results indicate there may be positive treatment effects across schools.

Multilevel Two-Tier Item Factor Model With Latent Change Parameterization

In item factor models, an item can potentially load on one or more latent dimensions (common factors). These common factors may also be potentially correlated, in the tradition of Thurston's (1947) multiple factor analysis. The bifactor model, a confirmatory item factor analysis model, has increasingly drawn interest among psychometricians. In a typical bifactor model, there is one primary dimension, representing a target construct being measured, and there are S specific dimensions that are also orthogonal conditionally on the general dimension, representing residual dependence above and beyond the general dimension. All items may load on the general dimension, and at the same time an item may load on at most one group-specific dimension.

Cai (2010b) proposed a two-tier item factor model for single-level data. It is minimally a more general version of the bifactor item factor model in which the number of general dimensions is not required to be equal to 1 and the correlations among these general dimensions may be explicitly represented and modeled. It may also be understood as a more general version of the Thurstonian correlated-factors multidimensional Item Response Theory (MIRT) model that explicitly includes an additional layer (tier) of random effects to account for residual dependence (e.g., due to repeated measures).

We employed a multilevel extension of the two-tier item factor analysis model as our measurement model. In a two-tier model for single-level data, two kinds of latent variables are specified, primary and group-specific. This creates a partitioning of the vector of latent variables ϑ_j for individual j into two mutually exclusive parts: $\vartheta_j = (\eta_j, \xi_j)$, where η_j is a vector of (potentially correlated) primary latent dimensions and ξ_j is a vector of specific dimensions that are independent conditional on the primary dimension. All latent variables in the model are random effects that vary over the individuals.

The MTT model, on the other hand, considers hierarchically nested data wherein individuals are nested within schools, for instance. In this model the latent variables for individual j in school k are partitioned into three mutually exclusive parts: $\vartheta_{jk} = (\theta_k, \eta_{jk}, \xi_{jk})$, where θ_k is the vector of level-2 (school-level) latent variables and η_{jk} and ξ_{jk} are the vectors of individual-level (level-1) primary and specific latent variables, respectively. The latent variables interact with item parameters to produce item response probabilities. For instance, a model that may work well with dichotomously scored item responses is the following extension of the classical 2-parameter logistic IRT model (see, e.g., Reckase, 2009):

$$P(Y_{ijk} = 1 | \vartheta_{jk}) = \frac{1}{1 + \exp[-(c_i + a_i' \vartheta_{jk})]}, \tag{2}$$

where Y_{ijk} is a Bernoulli random variable representing the response to item i from individual j in school k, c_i is the item intercept, and a_i is a conformable vector of item slopes. The model represents the response probability of a correct response ($Y_{ijk} = 1$) as a function of these item parameters and the latent variables. Obviously, $P(Y_{ijk} = 0 | \vartheta_{jk}) = 1.0 - P(Y_{ijk} = 1 | \vartheta_{jk})$. More complex item response models may be used depending on item types, such as the graded model for ordinal response data and the nominal categories model (see e.g., Cai, Yang, & Hansen, 2011). Continuous outcomes (e.g., conditional normal) or count outcomes (e.g., conditional Poisson) may be included.

Given $i = 1, \ldots, I$ items, $j = 1, \ldots, n_k$ individuals in school k, and $k = 1, \ldots, K$ schools, the observed data (marginal) likelihood function may take the following form:

$$L(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \int \prod_{j=1}^{n_k} \left[\iint \prod_{i=1}^{I} P(Y_{ijk} = y_{ijk} | \vartheta_{jk}) f(\xi_{jk}) d\xi_{jk} f(\eta_{jk}) d\eta_{jk} \right] f(\theta_k) d\theta_k \tag{3}$$

where y_{ijk} stands for the observed response to item i from individual j in school k, and $\boldsymbol{\gamma}$ stands for the collection of freely estimated model parameters. Yang, Monroe, and Cai (2012) developed efficient dimension reduction methods for maximum marginal likelihood estimation with the Bock-Aitkin (Bock & Aitkin, 1981) EM algorithm. Alternatively, the Metropolos-Hastings Robbins-Monro (MH-RM; Cai, 2010a, 2010b) algorithm may also be used to optimize the marginal likelihood function. Both algorithms are implemented in the flexMIRT software (Cai, 2013).

We applied the MTT model for calibrating and scaling pretest and posttest item responses in our multisite cluster randomized design. There are three key considerations in building the model for our study design and data structure. First, of the 23 items making up

the pretest and posttest outcome assessments, 20 items are in common. Thus, the design can be thought of as a test-retest administration of (essentially identical) assessments to the same group of individuals at two time points, before and after the intervention.

Cai (2010b) noted that when an IRT model must be calibrated with longitudinal item response data, even if the measurement instrument may be unidimensional at each time point, the multivariate longitudinal item data are inherently multidimensional. For designs with pretest and posttest, at least two occasion-specific primary latent dimensions are needed to model the initial status and potential gains in math knowledge, as well as to investigate potential differences in the structure of measurement (e.g., shifts in location or discrimination of items) over time, if necessary. In addition, the responses to the same item in pretest (time 1) and posttest (time 2) from the same individual may be residually correlated, even after controlling for the influence of the primary dimensions. Thus, item-specific residual correlation factors are introduced to handle the potential residual dependence, and there are as many of them as the number of repeated items.

Our randomized experiment consisted of two distinct groups—control and treatment—within each school. Thus, it is necessary to specify four within-school primary dimensions: pretest and posttest math knowledge variables in the control group and pretest and posttest math knowledge variables in the treatment group. This is akin to conditioning the latent math knowledge variables on the treatment assignment indicator variable, but this approach is more general because we allow both means and variances of latent variables to differ across treatment and control conditions, just as in a multiple-group model within each site. Finally, from the study design, it is clear that our data have a nested structure. Students (level-1 units) are nested within schools (level-2 units). More importantly, there are both control and treatment students within a school. Thus, both between-school and within-school variations in pretest and posttest latent variables need to be modeled.

Using conventional path diagrams, Figure 3 and Figure 4 show an exemplary multilevel two-tier item factor analysis model. Four pairs of common items are shown. We do not show all the items due to space constraints. The rectangles represent items, and circles represent latent variables. The four pairs of common items load on two between-school (level-2) primary dimensions, two within-school (level-1) primary dimensions, and four group-specific (level-1) dimensions in each condition.

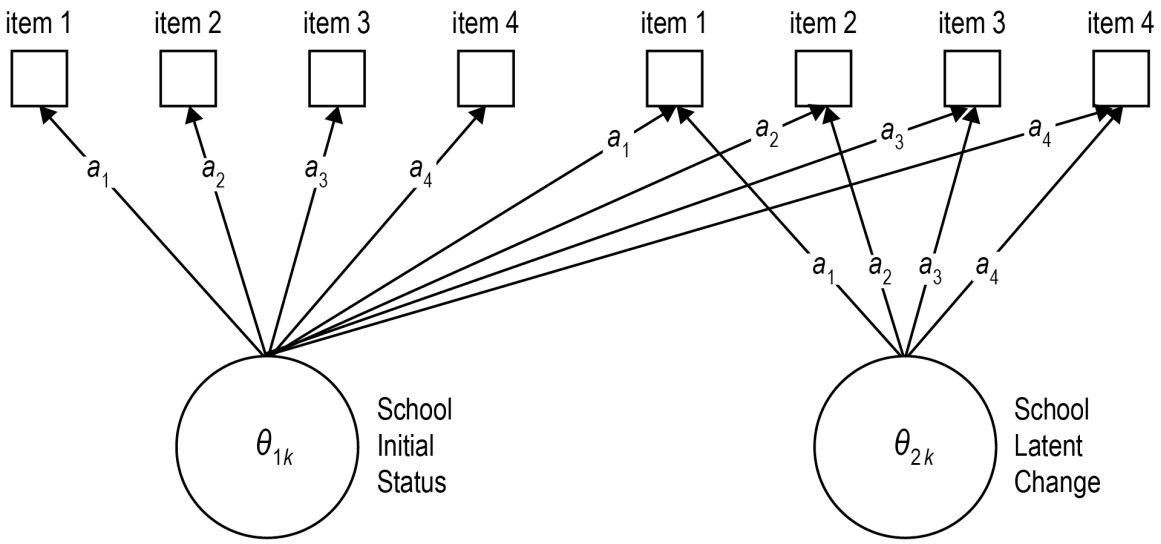

Figure 3. A multilevel two-tier item factor analysis model for multisite cluster randomized design with pretest and posttest design: Between-school model.

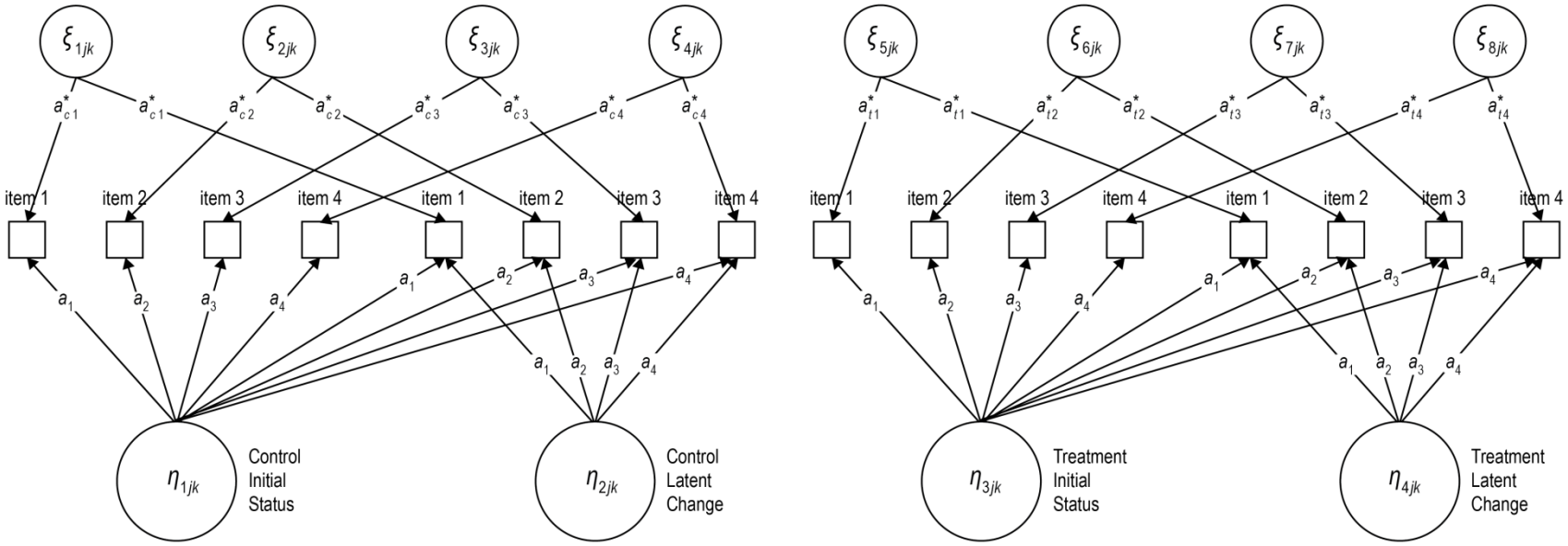

Figure 4. A multilevel two-tier item factor analysis model for multisite cluster randomized design with pretest and posttest design: Within-school model.

In terms of the factor pattern, the 16 × 14 factor pattern matrix corresponding to the model in Figure 4 has the following form:

$$\left(\begin{array}{cc|cc|cc|cccccccc}
a_1 & & a_1 & & & & a_{c1}^* & & & & & & & \\
a_2 & & a_2 & & & & & a_{c2}^* & & & & & & \\
a_3 & & a_3 & & & & & & a_{c3}^* & & & & & \\
a_4 & & a_4 & & & & & & & a_{c4}^* & & & & \\
a_1 & a_1 & a_1 & a_1 & & & a_{c1}^* & & & & & & & \\
a_2 & a_2 & a_2 & a_2 & & & & a_{c2}^* & & & & & & \\
a_3 & a_3 & a_3 & a_3 & & & & & a_{c3}^* & & & & & \\
a_4 & a_4 & a_4 & a_4 & & & & & & a_{c4}^* & & & & \\
a_1 & & & & a_1 & & & & & & a_{t1}^* & & & \\
a_2 & & & & a_2 & & & & & & & a_{t2}^* & & \\
a_3 & & & & a_3 & & & & & & & & a_{t3}^* & \\
a_4 & & & & a_4 & & & & & & & & & a_{t3}^* \\
a_1 & a_1 & & & a_1 & a_1 & & & & & a_{t1}^* & & & \\
a_2 & a_2 & & & a_2 & a_2 & & & & & & a_{t2}^* & & \\
a_3 & a_3 & & & a_3 & a_3 & & & & & & & a_{t3}^* & \\
a_4 & a_4 & & & a_4 & a_4 & & & & & & & & a_{t3}^*
\end{array}\right), \tag{4}$$

where nonempty entries indicate free parameters.

There are two between-school primary dimensions (item slopes in the first two columns above). The first dimension represents school k's initial status. The second dimension represents the potential posttest deviation from initial status for that school (e.g., due to exposure to business-as-usual math instruction outside of the game-based learning environment). For the control condition, the two within-school primary dimensions are in Columns 3 and 4. For the treatment condition, the two primary dimensions are in Columns 5 and 6. The interpretations of these dimensions resemble their between-school counterparts. Thus, there are a total of two between-school dimensions, four within-school primary dimensions, and eight within-school group-specific dimensions. Each residual dependence dimension is defined by an item pair within a condition. The constrained equal slope parameters on the item-specific residual dependence dimensions reflect an identification condition, as there is only one residual correlation per item pair. Note that Figure 3 and Figure 4 and the corresponding factor pattern matrix display only a part of model specification to the full data due to the space limitation. The size of the full factor pattern matrix is 90 (22 pretest plus 23 posttest items, times 2 for both conditions) × 46 (2 between, 4 within primary, and 40 within residual dependence dimensions), which is a truly high-dimensional model.

For a generic item i that appeared in both pretest and posttest, the linear predictor portions of the item response models can be written as the following:

$$\text{Pretest Control:} \quad a_i(\theta_{1k} + \eta_{1jk}) + a^*_{ci}\xi_{ijk} \tag{5-1}$$

$$\text{Posttest Control:} \quad a_i[(\theta_{1k} + \eta_{1jk}) + (\theta_{2k} + \eta_{2jk})] + a^*_{ci}\xi_{ijk} \tag{5-2}$$

$$\text{Pretest Treatment:} \quad a_i(\theta_{1k} + \eta_{3jk}) + a^*_{ti}\xi_{ijk} \tag{5-3}$$

$$\text{Posttest Treatment:} \quad a_i[(\theta_{1k} + \eta_{3jk}) + (\theta_{2k} + \eta_{4jk})] + a^*_{ti}\xi_{ijk} \tag{5-4}$$

It is seen that each item has an overall discrimination parameter a_i and the various latent variables contribute to the item response in a systematic manner. Specifically, two between-school latent variables, θ_{1k} and θ_{2k}, represent latent initial status and latent gain between pretest and posttest for school *k*, respectively. In addition, among the four within-school latent variables, the first two latent variables represent initial status (η_{1jk}) and latent gain (η_{2jk}) for student *j* in the control condition within school *k*, and the rest represent initial status (η_{3jk}) and latent gain (η_{4jk}) for student *j* in the treatment condition within school *k*. As such, the additional latent variables at posttest represent potential gains over the pretest level. Furthermore, the variation is decomposed at both pretest and posttest into between-school and within-school components. By allowing mean differences between η_2 and η_4 to be estimated, that is, the difference between latent changes between the treatment and control conditions within schools, potential effects of treatment on learning gain may be explicitly represented. This model is motivated by growth modeling developments as represented in Bock and Bargmann (1966), Embretson (1991), Cai (2010b), and McArdle (2009), among others. Upon estimating the item parameters, IRT scaled scores can be computed for each of the latent variables as posterior means or as multiple imputations (plausible values).

There are several new features of the MTT model that are particularly relevant to treatment impact evaluations. First, we estimate mean and variance of pretest and posttest latent variables separately for the treatment and control groups. This approach prevents us from the inadvertent bias induced by shrinking the individual posteriors of both treatment and control conditions to a common mean, when the treatment is expected to differentially impact the mean and variance in each condition. It is a standard practice in IRT modeling for off-the-shelf assessments that no information beyond the test items themselves (plus distributional assumptions about the latent variable, usually presumed standard normal) is used in estimating the scaled scores. Thus virtually all off-the-shelf IRT analyses and scaled outcome scores assume full exchangeability of the treatment and control students. While perhaps needed for practical and legal reasons in summative assessments, we argue that in an experimental context, the failure to include important conditioning information (e.g.,

treatment assignment) in the measurement model will lead to inconsistencies in the subsequent inference about treatment effect. It is interesting to note that this is not an entirely new observation. In the context of large-scale educational surveys (e.g., NAEP), researchers have long argued for the importance of including population conditioning covariate information into the measurement model so that estimates from student survey data are statistically consistent for the population comparisons of interest (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992). Similarly, when we adopt the view that all latent variables can be viewed as missing data, we could not help but notice that this is exactly the same argument made in the multiple imputation literature for survey nonresponse (e.g., Little & Rubin, 2002) that important conditioning information from the data analysts' model must be present in the imputation model for the imputation-based inferences to be valid. We take the conditional exchangeability one step further by including a range of observed and latent variables to make the measurement model more commensurate with the research design.

Second, in the IRT model calibration stage, latent gain/change scores are represented explicitly in the MTT model. This approach is particularly effective when examining change from two-wave (pre-post) data, in that the variance of treatment latent gain scores is reduced by an amount proportional to the shared variance between pretest and posttest. In other words, the model utilizes the availability of a stable and randomly equivalent control condition where there should be (theoretically) only normative change in student posttest performance from pretest to decompose the observed variability in the outcome assessment into two components that may be provisionally termed latent prior knowledge factor and latent malleable factor. As soon as the model-based decomposition is achieved, we submit that the comparison between treatment and control conditions should be focused on the latent malleable factors. This is in stark contrast with the standard approach, whether summed score based or IRT based, in which the posttest outcome score is used directly for impact estimation. We argue that the standard approach is less optimal because the variation in the posttest outcome (observed or scaled latent variable estimates) conflates two sources of variance, that is, that of prior knowledge and that of malleable difference. One direct consequence of our approach is that the resulting reduced variance in latent gain estimates may lead to substantially increased effect size of the treatment effect (to be elaborated).

Figure 5 (also see Table A3 in Appendix A) displays descriptive statistics of the estimated latent pretest score and latent change score by experimental conditions. The means of latent pretest scores for both groups are very nearly 0 and the standard deviations are also close to 1. However, the latent change score means for the control and treatment groups are,

respectively, 0.022 and 0.341. Thus, the observed difference is equal to 0.319. The outstanding pattern is that the standard deviation of the latent change score is far smaller than the standard deviation of latent pretest score. The standard deviation of the latent change score for the control is 0.335 and for the treatment is 0.448. The pooled standard deviation is about 0.40, which is only 40% of the latent pretest score's standard deviation.

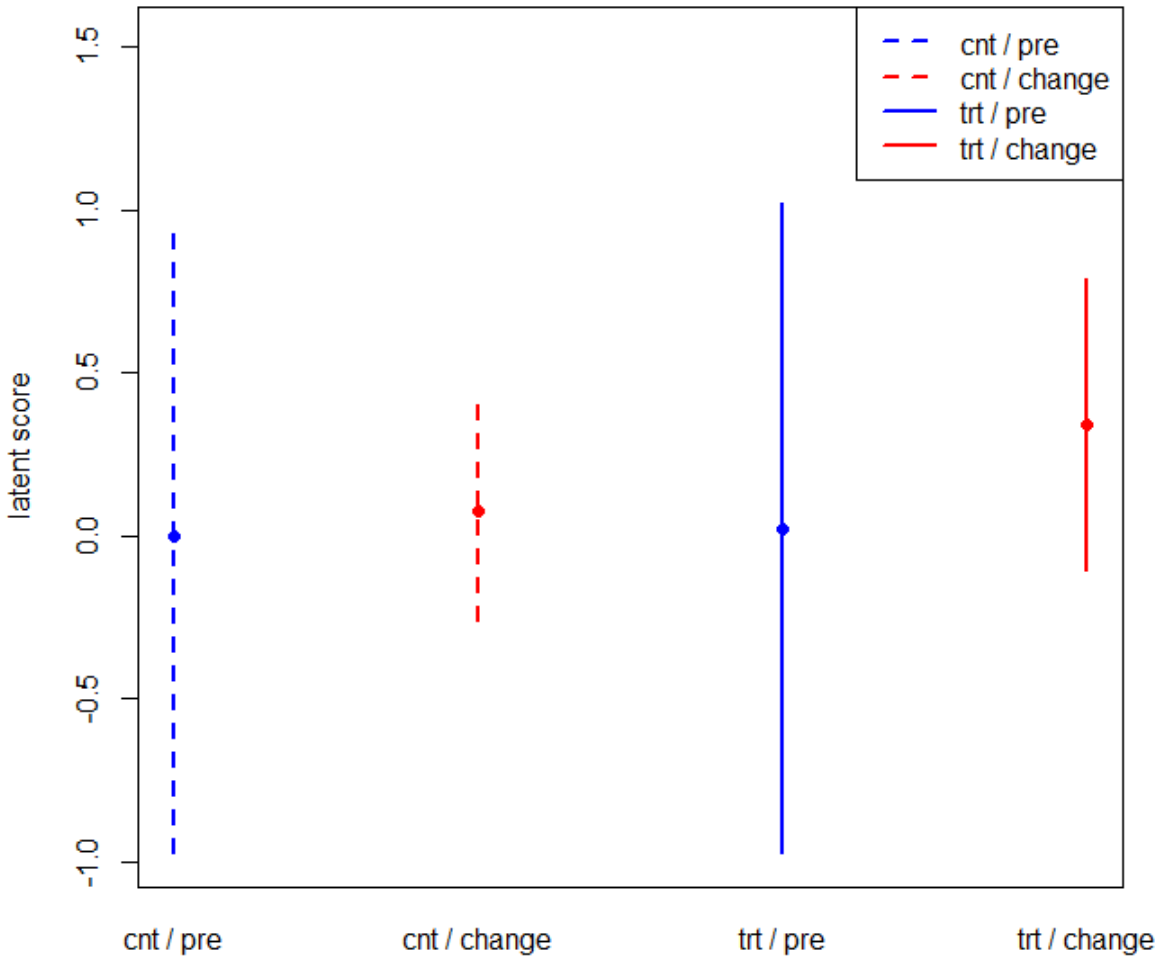

Figure 5. Descriptive statistics of latent pretest scores and latent change scores by experimental conditions.

This reduced variance of latent change score is necessitated by the fact that the latent posttest score can be viewed as a sum of initial status and latent gain (Figure 6). In other words, if we treat the total variance of latent posttest as constant, the latent change score is equal to the latent pretest score minus the latent pretest score, so the variance has to be smaller. When we view the gain score variance in the control condition as representing the variability of the malleable factor distribution in the population of interest, it becomes the more natural unit of comparison for computing standardized effect sizes.

Finally, we note that the particular factor pattern of the MTT model (both between and within schools) lends itself to another interpretation that may be useful to some. The latent gain score dimensions are effectively a specific dimension (as in a bifactor model) if the latent initial status dimension is regarded as the primary dimension, albeit with additional equality constrains. Adopting the standard bifactor or hierarchical item factor model interpretation, the specific dimensions represent residualized variation above and beyond the pretest variation (Reise, 2012). As such, they isolate that part of the posttest individual differences in performance that is specific, after controlling for prior knowledge. They are more sensitive measures for targeted interventions than the observed outcome scores (see e.g., Gibbons et al., 2008, p. 365).

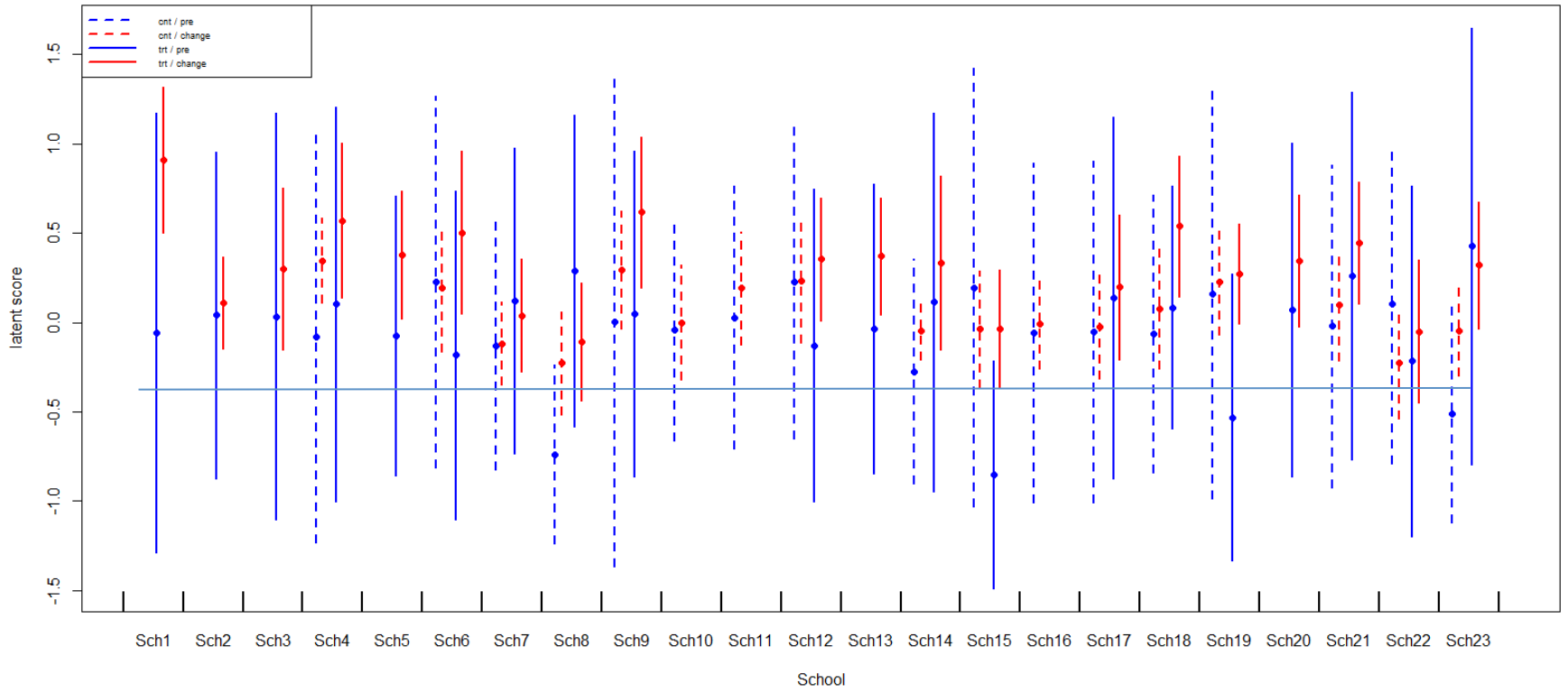

Figure 6. Descriptive statistics of latent pretest score and latent change score by schools and experimental conditions.

We also plot the means plus/minus one standard deviation for the latent initial status and latent gain scores for each condition within a school. The most noticeable effect is that both treatment and control groups in almost all the schools now show positive change. Furthermore, such positive change is larger for the treatment condition than for the control condition. In addition, the standard deviation of the latent change score is significantly smaller for the standard deviation of the latent pretest score. As we see in Figure 5, the variance in the latent change score is about 40% of the variance in the latent pretest score.

Impact Model: Two-Level Hierarchical Model

We employed a standard two-level hierarchical model to estimate the treatment effect in multisite cluster randomized trial design. The following level-1 (within-school) model in Equation 6 specifies a model with outcome variable, Y_{ij}, the raw total posttest score, for student i in school j, as a function of treatment indicator variable, Trt_{ij}. Note that the treatment indicator variable takes a value of -0.5 for the control condition, 0.5 for the treatment condition within each school j. By virtue of this coding, β_{0j} represents the mean posttest score for school j and β_{1j} represents the expected difference in the outcome between treatment and control conditions in school j.

$$Y_{ij} = \beta_{0j} + \beta_{1j} * Trt_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{6}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad u_{0j} \sim N(0, \tau_{00}) \tag{7-1}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad u_{1j} \sim N(0, \tau_{11}) \tag{7-2}$$

The level-1 error term, ε_{ij}, is assumed to be normally distributed with mean 0 and variance, σ^2. At level 2, β_{0j} and β_{1j} are modeled as a function of the grand means and the random effects around the means, u_{0j} and u_{1j} respectively. The regression coefficient γ_{10} represents the overall treatment effect. The two random effects are assumed to be bivariate normal with variances τ_{00} and τ_{11}, and covariance τ_{01}.

In addition, we specify another two-level HLM which includes observed pretest summed score as a covariate. This variable is group-mean centered so β_{0j} still represents the mean posttest score for school j. The key parameter of interest, β_{1j}, becomes the expected difference in posttest score between the two groups in school j, holding pretest constant.

$$Y_{ij} = \beta_{0j} + \beta_{1j} * Trt_{ij} + \beta_{2j} * Pretest_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{8}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}) \tag{9-1}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11}) \tag{9-2}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}, \quad u_{2j} \sim N(0, \tau_{22}) \tag{9-3}$$

Impact model results using raw total scores. Using those two models above, we analyzed two data sets: (1) total raw pretest and posttest scores, and (2) MTT scaled score.

The results from total raw scores are presented in Table 2. The grand mean of raw posttest total score posttest (γ_{00}) is approximately 10.1 and the expected overall difference between treatment and control (γ_{10}) is 1.4, which are all statistically significant. This expected difference, 1.4, is about 0.23 of pooled standard deviation of posttest raw score. Also, the variance of the treatment effect across schools is 5.7 and its 95% interval of the expected differences across schools ranges from -3.28 to 6.05. Note, however, that this variability does not take the pretest differences into account.

Table 2

Impact Model Result: Raw Total Pretest and Posttest

	Model 1: Unconditional			Model 2: Pretest as covariate		
Fixed effects	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Intercept (γ_{00})	10.062	0.489	< .0001	10.334	0.486	< .0001
Trt (γ_{10})	1.388	0.636	.029	1.119	0.332	.001
Pretest(γ_{20})				0.9304	0.026	< .0001
Variance components	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Level-1 (σ^2)	19.914	0.728	< .0001	7.919	0.298	< .0001
Intercept (τ_{00})	4.723	1.634	.002	5.178	1.646	.001
Trt (τ_{11})	5.669	2.436	.010	1.307	0.686	.029
Pretest (τ_{22})				0.005	0.298	< .0001

In the second panel of Table 2, we present the result from the model where pretest score is included as a covariate as specified in Equations 8, 9-1, 9-2, and 9-3. As can be seen, the pretest score is positively associated with the posttest (i.e., one unit change of pretest leads to a 0.93 increase in posttest). After controlling for pretest difference, the overall treatment effect is approximately 1.1 and its variance is 1.3, which becomes much smaller than the corresponding variance in the previous unconditional model. The lower and upper ends of the 95% interval of the treatment effects across schools are -1.05 and 3.43, respectively.

Figure 7 shows each school's empirical Bayes (EB) estimate of treatment effect which is obtained from Model 2. The middle bar in each school (x-axis) represents the EB estimate and the vertical line represents its 95% interval. Also, the solid horizontal line represents the overall treatment effect ($\gamma_{10} = 1.119$), whereas the dotted horizontal line is the reference

line whether each school's treatment effect is statistically significantly different from zero. Ten out of the 23 schools show intervals that do not cover 0, which means that students in these schools performed statistically significantly better in the treatment than those in the control conditions. Nine of the remaining 11 schools (the exceptions are Schools 15 and 19) have positive EB estimate, but their treatment effects are not statistically significant as their 95% intervals cover a value of 0.

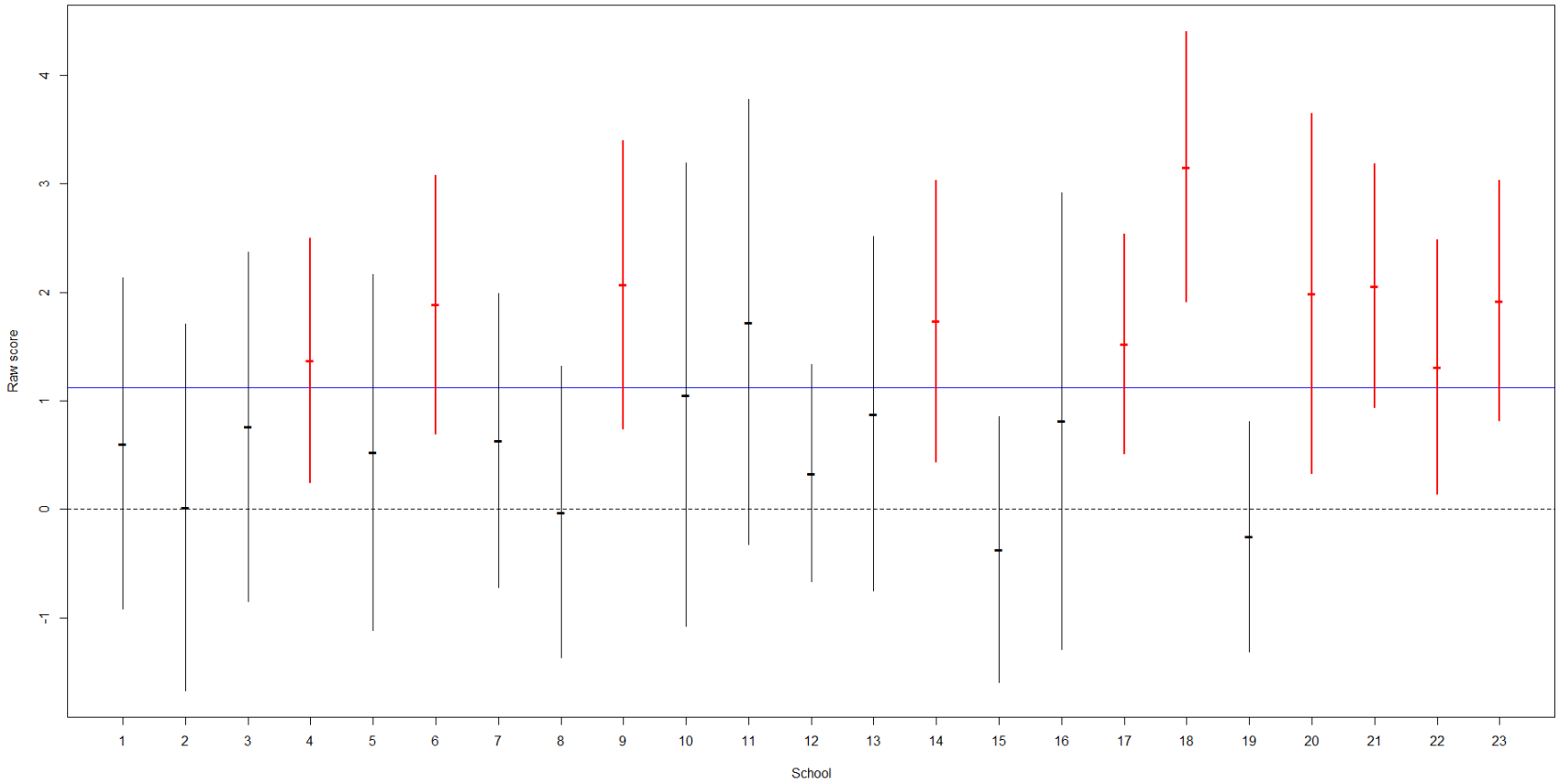

Figure 7. Empirical Bayes estimate of treatment effect by schools: Conditional model result with raw total score.

Impact model results using MTT scaled scores. We also fit the previously specified multilevel models 1 and 2 to the latent gain scaled scores obtained from the MTT model. Note that the outcome of the analysis is the latent change scores instead of the posttest score. As can be seen in Table 3, the overall difference in the latent change score between the treatment and control group is 0.243 and its variance is 0.013. The resulting 95% interval estimate of the differences between the two conditions across schools ranges from 0.466 to 0.021. Interestingly, the overall treatment effect does not change much even after controlling for latent pretest score in the model, as we argued that the MTT model has already partialed out the initial status difference. The coefficient (γ_{10}) in Model 2 is 0.236 and its variance reduced to 0.008. The overall treatment effect is approximately 0.58 standard deviation of the latent change score.

Table 3

Impact Model Result: Latent Gain Scores from the MTT Model

	Model 1: Unconditional			Model 2: Pretest as covariate		
Fixed effects	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Intercept (γ_{00})	0.199	0.042	.0001	0.207	0.042	< .0001
Trt (γ_{10})	0.243	0.035	< .0001	0.236	0.029	< .0001
Pretest(γ_{20})				0.107	0.009	< .0001
Variance components	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Level-1 (σ^2)	0.119	0.012	< .0001	0.110	0.004	< .0001
Intercept (τ_{00})	0.038	0.008	.001	0.037	0.012	.001
Trt (τ_{11})	0.013	0.004	.049	0.008	0.006	.094
Pretest (τ_{22})				0.000	0.298	.423

Finally, we present EB estimate of the treatment effects and the 95% confidence intervals based on Model 2 for each school in Figure 8. The horizontal line represents the overall treatment effect, which is 0.236. The most noteworthy thing in this figure is that 22 schools' lower 95% confidence limit is above zero. This indicates that the treatment effects in 22 schools except for one school (School 8) are statistically significant.

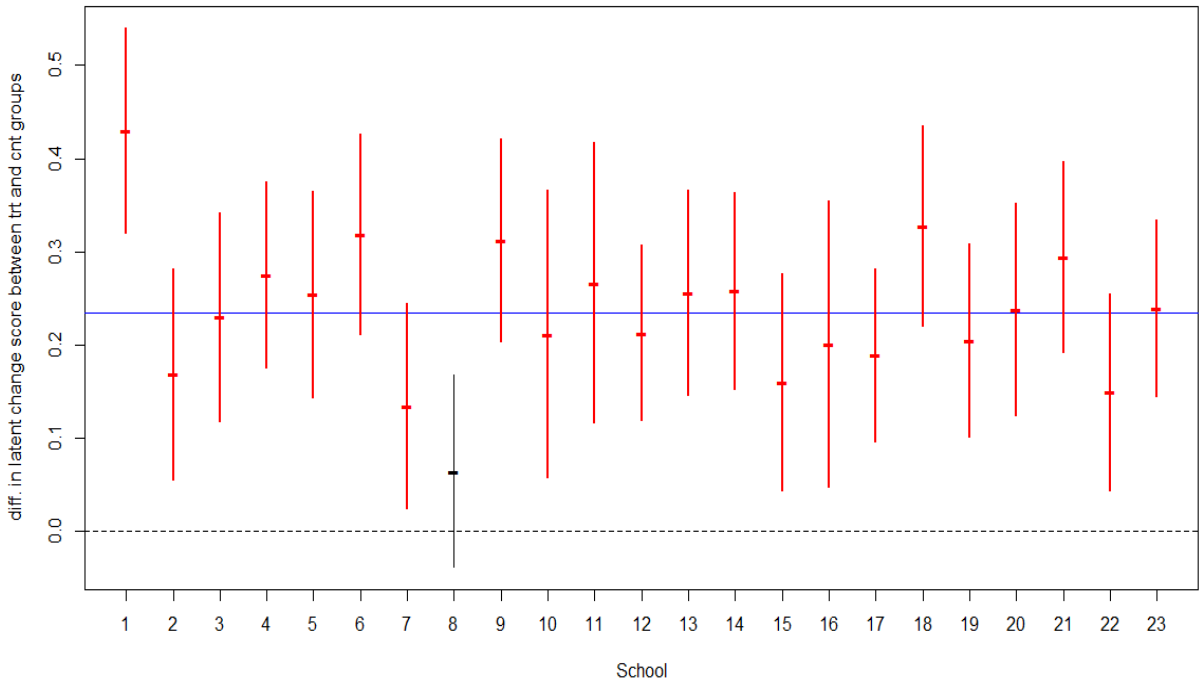

Figure 8. Empirical Bayes estimate of treatment effect by schools: Difference in latent change score between the treatment and control groups.

Effect sizes. Effect sizes (ES) are reported and used extensively in randomized trials as they provide more meaningful standardized interpretations of treatment effect estimates. We calculated ES of the treatment effect obtained from multilevel impact Model 2. Three different effect sizes were calculated as follows:

Cohen's δ (Cohen, 1992) = γ_{10} / SD of outcome

Hedges's ES (Hedges & Rhoads, 2010) = γ_{10} / sqrt ($\sigma^2 + \tau_{00} + \tau_{11}$)

Conditional ES (Spybrook, Raudenbush, Congdon, & Martinez, 2009) = γ_{10} / sqrt (σ^2)

Cohen's δ is colloquially referred to as "the ES" and is calculated by dividing the treatment effect coefficient by the standard deviation of the outcome. Thus, the treatment effect is phrased in terms of the standard deviations of the outcome measure. Hedges's ES is comparable to Cohen's δ except for the fact that it uses a model-based standard deviation of the outcome. In other words, the square root of the sum of all the variance components in an unconditional two-level hierarchical model for the multisite randomized trials is placed in the denominator instead of the standard deviation of outcome in the case of Cohen's δ. Lastly, Spybrook et al. (2009) introduced the conditional ES in their Optimal Design software, which is by far the most popular software program for conducting statistical power analysis for various cluster randomized design studies. Conditional ES uses only level-1 variance

which is reduced after pretest score being included so it is generally bigger than either Cohen's δ or Hedges's ES.

Table 4

Effect Sizes

Outcome scores	Est. (γ_{10})	Posttest *SD*	Lev-1 var(σ^2)	Lev-2 Int var(τ_{00})	Lev-2 Trt var(τ_{11})	Hedges's ES	Cond. ES	Cohen's δ
Raw/summed	1.139	4.986	19.914	4.723	5.669	0.207	0.405	0.228
Latent change	0.236	0.419	0.110	0.037	0.009	0.615	0.705	0.579

As we see in Table 4, the raw score model yielded an ES of roughly 0.2. Although 0.2 ES is viewed as a small effect size, given the very short duration and low frequencies of treatment, it should be considered as a fairly significant effect. Furthermore, our MTT latent change score model produced an ES in excess of 0.55. In addition to the results in Figure 8 that display statistically significant treatment effects for 22 out of 23 schools in the sample as compared to 10 out of 23 schools in the raw score model, the MTT latent change score model also significantly improved the ES estimate compared to the raw score model ES.

Summary

The instructional effects of video games developed by CATS were examined with a large-scale randomized trial with over 1500 students in 30 treatment classrooms and 29 control classrooms in 26 schools in 9 districts. The video games were intended to improve students' mathematics learning outcomes as measured by items similar to pre-algebra mathematics standardized assessment items on rational numbers and fractions. Students in treatment classrooms played games on the topic of rational numbers and fractions, whereas those in the control classrooms played an alternative set of games on solving equations.

When data were analyzed with the standard approach using posttest summed scores as the outcome variable, results indicated a small (but positive and statistically significant) effect size (.23 Cohen's *d*). When the outcome variable was constructed from a multilevel multidimensional latent variable modeling approach, the effect size improved to medium to large range (approximately .6). The new latent variable modeling based outcome measure was more sensitive to instructional intervention than standard measurement approaches.

This study not only demonstrated the effectiveness of carefully designed learning games on student outcomes, but also proposed a generalizable solution to measurement error and multilevel modeling issues in multisite randomized trials. We argued that the traditional

measurement approaches using either raw summed scores or "off-the-shelf" IRT-based scaled scores ignored the plausibility of certain inherent exchangeability assumptions. Appropriately integrating measurement modeling with treatment impact modeling in multisite randomized studies with repeated measures requires the careful specification of model features that serve to explain (1) the dependency between the latent outcome variables at each occasion; (2) item-level residual dependence due to repeated measures; (3) lack of full exchangeability of participants between treatment and control conditions; and (4) individual nesting within sites. In contrast, we illustrated how the multilevel two-tier item factor model may be used to address each of the four aspects mentioned above to arrive at results that may deserve attention from both substantive and methodological perspectives.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Bargmann, R.E. (1966). Analysis of covariance structures. *Psychometrika, 31,* 507–533.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75,* 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35,* 307–335.

Cai, L. (2013). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221–248.

Center for Advanced Technology in Schools (CATS). (2012). *CATS Developed Games* (CRESST Resource Report No. 15). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Curran, P., Bauer, D., & Willoughby, M. (2004). Testing main effects and interactions in latent curve analysis. *Psychological Methods, 9,* 220–237.

Curran, P., & Bollen, K. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 105–136). Washington, DC: American Psychological Association.

Curran, P., Hussong, A., Cai, L., Huang, W., Chassin, L., Sher, K., & Zuchek, R. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44*(2), 365–380.

de Freitas, S. (2006). *Learning in immersive worlds: A review of game-based learning.* London: Joint Information Systems Committee (JISC).

Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research & Development, 53*(2), 67–83.

Donchin, E. (1989). The learning strategies project. *Acta Psychologica, 71,* 1–15.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave/Macmillan.

Gee, J. P. (2004). Learning by design: Games as learning machines. *Interactive Educational Multimedia, 8,* 15–23.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361–368.

Hedges, L., & Rhoads, C. (2010). *Statistical Power Analysis in Education Research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Kafai, Y. B. (2006). Playing and making games for learning: Instructionist and constructionist perspectives for game studies. *Games and Culture, 1*, 36–40.

Kafai, Y. B., Franke, M., Ching, C., & Shih, J. (1998). Game design as an interactive learning environment fostering students' and teachers' mathematical inquiry. *International Journal of Computers for Mathematical Learning, 3*, 149–184.

Kirriemuir, J., & McFarlane, A. E. (2003). *Literature review in games and learning*, Report 8. Bristol: Nesta Futurelab.

Klopfer, E., & Squire, K. (2004). Getting your socks wet: Augmented reality environmental science. In *Proceedings of the 6th international conference on the learning sciences (ICLS)* (p. 614), Los Angeles, CA. Retrieved October 29, 2007 from http://portal.acm.org/citation.cfm&id-1149238.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society - Series B (Methodological), 34,* 1–41.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 4,* 333–369.

Malone, T. W., & Lepper, M. R. (1987). *Aptitude, learning and instruction III: Cognitive and affective process analysis.* Hillsdale, NJ: Erlbaum.

Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 281–305). Charlotte, NC: Information Age.

McArdle, J. J. (2009). Latent variable modeling of difference and changes with longitudinal data. *Annual Review of Psychology, 60,* 577–605.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B. K., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.

National Mathematics Advisory Panel (NMAP). (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

O'Neil, H. F., & Perez, R. S. (2008). *Computer games and team and individual learning*. Oxford, UK: Elsevier.

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal, 16*(4), 455–474.

Ramsberger, P. F., Hopwood, D., Hargan, C. S., & Underhill, W. G. (1983). *Evaluation of a spatial data management system for basic skills education. Final Phase I Report for Period 7 October 1980- 30 April 1983* (HumRROFR-PRD-83-23). Alexandria, VA: Human Resources Research Organization.

Reckase, M. (2009). *Multidimensional item response theory.* New York: Springer.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.

Ruben, B. D. (1999). Simulations, games, and experience-based learning: The quest for a new paradigm for teaching and learning. *Simulation & Gaming, 30,* 498–505.

Spybrook, J., Raudenbush, S.W., Congdon, R., & Martinez, A. (2009). Optimal Design for longitudinal and multilevel research (Version 2.0) [Computer software and documentation]. Available at www.wtgrantfoundation.org

Squire, K. (2011). *Video games and learning: Teaching and participatory culture in the digital age*. New York, NY: Teachers College Pres.

Thomas, P., & Macredie, R. (1994). Games and the design of human-computer interfaces. *Educational Technology, 31*(2), 134–142.

Thurston, L.L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Tobias, S., Fletcher, J. D., Bediou, B., Wind, A. P., & Chen, F. (2014). Multimedia learning from computer games. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 762–784). New York, NY: Cambridge University Press.

Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127–222). Charlotte, NC: Information Age.

Tobias, S., Fletcher, J. D., & Wind, A. (2014). Game based learning. In M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational and Communications Technology* (4th ed., pp. 485–504). New York: Springer Academic.

Vendlinski, T. P., Delacruz, G. C., Buschang, R. E., Chung, G. K. W. K., & Baker, E. L. (2010). *Developing high-quality assessments that align with instructional video games* (CRESST Rep. No. 774). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wiener, N. (1954). *The human use of human beings: Cybernetics and society*. New York, NY: Da Capo Press.

Yang, J., Monroe, S., & Cai, L. (2012). *A multiple group multilevel item bifactor analysis model.* Paper presented at 2012 Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullen, G., Lai, B., …Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research, 82*, 61–89. doi:10.3102/003465431243698010.3102/0034654312436980.

Appendix A:
Descriptive Statistics of Pretest and Posttest Scores by Schools and Conditions

Table A1

Descriptive Statistics of Raw Total Scores of Pretest and Posttest by Conditions

Condition	Variable	*n*	Mean	*SD*	Min.	Max.
Control	Pretest	763	7.97	3.99	0	22.0
	Posttest	745	9.47	4.58	0.5	22.5
Treatment	Pretest	808	8.04	4.13	0.25	21.0
	Posttest	792	10.91	5.22	1	22.9

Table A2

Descriptive Statistics of Raw Total Scores of Pretest and Posttest by Schools and Conditions

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
1	Treatment	Pretest	64	6.86	4.40	0.83	20.00
		Posttest	57	11.37	5.86	2.00	22.88
2	Treatment	Pretest	32	5.32	2.85	1.00	10.83
		Posttest	32	6.63	3.62	1.83	16.54
3	Treatment	Pretest	34	9.13	4.27	2.00	17.42
		Posttest	33	12.08	6.13	2.38	22.54
4	Control	Pretest	31	7.78	4.37	0.00	15.17
		Posttest	31	10.72	4.53	3.17	17.79
	Treatment	Pretest	53	8.78	4.60	1.00	21.00
		Posttest	52	12.59	4.78	4.00	22.50
5	Treatment	Pretest	44	6.15	2.99	1.00	14.25
		Posttest	43	9.29	4.28	2.00	19.83
6	Control	Pretest	32	9.27	4.17	2.33	18.00
		Posttest	29	10.84	4.89	1.33	21.75
	Treatment	Pretest	30	7.41	3.44	0.83	17.25
		Posttest	29	11.18	5.01	1.17	20.88
7	Control	Pretest	23	5.01	2.63	0.00	10.92
		Posttest	24	5.82	2.08	2.00	9.08
	Treatment	Pretest	25	6.38	3.13	1.33	12.75
		Posttest	26	7.72	3.65	2.00	14.41
8	Control	Pretest	17	3.92	1.77	1.17	7.50
		Posttest	19	5.35	2.50	2.00	10.63
	Treatment	Pretest	48	7.61	3.41	2.50	14.92
		Posttest	49	8.23	3.83	1.00	15.91
9	Control	Pretest	25	12.89	5.23	3.00	22.00
		Posttest	18	14.11	5.21	4.88	21.38
	Treatment	Pretest	26	13.13	3.83	4.83	18.58
		Posttest	26	17.21	4.38	8.50	22.13
10	Control	Pretest	26	7.88	2.15	4.50	12.00
		Posttest	24	9.34	3.82	3.55	18.63
11	Control	Pretest	58	9.58	3.07	4.00	16.67
		Posttest	55	11.65	3.82	3.00	17.96

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
12	Control	Pretest	57	8.34	3.46	1.00	15.50
		Posttest	57	10.82	4.54	3.38	20.96
	Treatment	Pretest	64	7.62	3.28	1.00	16.58
		Posttest	63	10.31	4.39	2.75	20.13
13	Treatment	Pretest	54	7.82	3.10	2.17	14.75
		Posttest	51	10.66	3.93	2.83	18.83
14	Control	Pretest	16	6.38	2.90	3.00	11.92
		Posttest	16	7.29	1.91	4.08	10.80
	Treatment	Pretest	37	8.07	4.15	1.67	16.33
		Posttest	42	10.61	5.34	2.25	20.38
15	Control	Pretest	77	8.45	4.60	1.00	22.00
		Posttest	79	9.82	5.20	3.00	22.38
	Treatment	Pretest	16	5.12	2.23	2.83	10.00
		Posttest	15	5.19	2.52	1.17	11.16
16	Control	Pretest	28	6.39	3.40	0.00	14.16
		Posttest	25	7.53	3.58	2.17	16.63
17	Control	Pretest	65	9.14	3.91	1.00	17.59
		Posttest	66	10.38	4.52	3.13	20.63
	Treatment	Pretest	64	9.98	3.90	3.00	19.00
		Posttest	63	13.07	5.08	2.50	21.88
18	Control	Pretest	25	9.81	3.65	3.00	21.00
		Posttest	21	11.17	4.10	4.83	18.79
	Treatment	Pretest	29	9.64	3.19	3.08	15.75
		Posttest	29	14.49	3.60	7.25	20.13
19	Control	Pretest	102	7.23	4.22	1.00	19.17
		Posttest	99	9.32	5.07	2.00	22.54
	Treatment	Pretest	31	4.77	2.75	1.00	12.83
		Posttest	28	6.03	2.31	2.00	10.67
20	Treatment	Pretest	26	11.22	3.69	1.33	17.33
		Posttest	25	14.58	4.72	2.33	20.58
21	Control	Pretest	93	7.75	3.43	1.33	16.33
		Posttest	92	9.15	4.23	0.50	20.13
	Treatment	Pretest	32	8.90	4.34	0.25	20.00
		Posttest	32	12.08	4.69	3.55	22.88

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
22	Control	Pretest	37	7.65	3.79	1.33	18.42
		Posttest	36	8.11	3.46	0.92	18.63
	Treatment	Pretest	35	6.35	3.47	0.50	15.42
		Posttest	30	8.40	4.18	3.25	18.12
23	Control	Pretest	51	5.77	2.03	1.50	10.92
		Posttest	54	6.88	2.89	2.58	15.71
	Treatment	Pretest	64	9.36	4.76	1.75	19.67
		Posttest	67	12.22	5.13	2.83	21.88

Table A3

Descriptive Statistics of MTT Scale Scores of Pretest and Posttest by Conditions

Conditions	Variable	*n*	Mean	*SD*	Min.	Max.
Control	Pretest	763	-0.002	0.975	-2.433	3.471
	Posttest	745	0.073	0.335	-0.917	1.215
Treatment	Pretest	808	0.022	0.999	-2.736	3.555
	Posttest	792	0.341	0.448	-0.884	1.877

Table A4

Descriptive Statistics of MTT Scaled Scores of Pretest and Posttest by Schools and Conditions

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
1	Treatment	Pretest	64	-0.060	1.230	-2.248	3.407
		Posttest	57	0.909	0.407	0.029	1.835
2	Treatment	Pretest	32	0.041	0.915	-1.523	1.806
		Posttest	32	0.111	0.258	-0.489	0.597
3	Treatment	Pretest	34	0.033	1.137	-1.962	2.196
		Posttest	33	0.298	0.453	-0.632	1.247
4	Control	Pretest	31	-0.081	1.152	-2.069	1.767
		Posttest	31	0.346	0.236	-0.302	0.738
	Treatment	Pretest	53	0.101	1.106	-1.979	3.249
		Posttest	52	0.570	0.435	-0.207	1.877
5	Treatment	Pretest	44	-0.076	0.784	-1.698	1.989
		Posttest	43	0.378	0.358	-0.624	0.946
6	Control	Pretest	32	0.226	1.041	-1.569	2.718
		Posttest	29	0.191	0.358	-0.505	0.980
	Treatment	Pretest	30	-0.182	0.920	-2.088	2.121
		Posttest	29	0.502	0.457	-0.470	1.599
7	Control	Pretest	23	-0.130	0.697	-1.485	1.354
		Posttest	24	-0.118	0.232	-0.571	0.441
	Treatment	Pretest	25	0.121	0.855	-1.506	1.614
		Posttest	26	0.037	0.316	-0.612	0.632
8	Control	Pretest	17	-0.740	0.501	-1.553	0.137
		Posttest	19	-0.229	0.289	-0.642	0.437
	Treatment	Pretest	48	0.287	0.872	-1.518	1.916
		Posttest	49	-0.108	0.330	-0.784	0.569
9	Control	Pretest	25	0.002	1.369	-2.433	2.413
		Posttest	18	0.294	0.329	-0.172	0.759
	Treatment	Pretest	26	0.047	0.912	-1.959	1.338
		Posttest	26	0.617	0.423	-0.118	1.474
10	Control	Pretest	26	-0.043	0.618	-1.107	1.236
		Posttest	24	0.000	0.322	-0.474	0.957
11	Control	Pretest	58	0.028	0.733	-1.753	1.252
		Posttest	55	0.191	0.317	-0.674	0.807

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
12	Control	Pretest	57	0.224	0.877	-1.166	1.995
		Posttest	57	0.234	0.351	-0.620	0.971
	Treatment	Pretest	64	-0.129	0.874	-2.036	1.829
		Posttest	63	0.355	0.344	-0.621	1.016
13	Treatment	Pretest	54	-0.038	0.811	-1.547	1.816
		Posttest	51	0.370	0.326	-0.521	1.048
14	Control	Pretest	16	-0.274	0.631	-1.330	1.006
		Posttest	16	-0.050	0.162	-0.408	0.201
	Treatment	Pretest	37	0.113	1.060	-1.701	2.136
		Posttest	42	0.331	0.487	-0.630	1.477
15	Control	Pretest	77	0.194	1.227	-2.058	3.471
		Posttest	79	-0.039	0.326	-0.828	1.215
	Treatment	Pretest	16	-0.853	0.638	-1.614	0.684
		Posttest	15	-0.036	0.332	-0.490	0.759
16	Control	Pretest	28	-0.059	0.951	-1.988	1.901
		Posttest	25	-0.008	0.250	-0.490	0.493
17	Control	Pretest	65	-0.055	0.956	-2.109	2.027
		Posttest	66	-0.025	0.292	-0.681	0.645
	Treatment	Pretest	64	0.136	1.013	-1.910	2.171
		Posttest	63	0.196	0.403	-0.749	1.062
18	Control	Pretest	25	-0.063	0.778	-1.231	2.256
		Posttest	21	0.075	0.334	-0.394	0.768
	Treatment	Pretest	29	0.084	0.679	-1.363	1.121
		Posttest	29	0.538	0.395	-0.261	1.336
19	Control	Pretest	102	0.160	1.147	-2.042	2.920
		Posttest	99	0.224	0.293	-0.514	0.922
	Treatment	Pretest	31	-0.532	0.802	-1.942	1.536
		Posttest	28	0.270	0.281	-0.459	0.850
20	Treatment	Pretest	26	0.070	0.933	-2.736	1.894
		Posttest	25	0.342	0.369	-0.376	0.846
21	Control	Pretest	93	-0.021	0.903	-1.960	1.691
		Posttest	92	0.097	0.314	-0.718	1.055
	Treatment	Pretest	32	0.259	1.029	-1.954	3.555
		Posttest	32	0.445	0.340	-0.052	1.369

School	Condition	Variable	*n*	Mean	*SD*	Min.	Max.
22	Control	Pretest	37	0.105	0.896	-1.858	2.500
		Posttest	36	-0.225	0.313	-0.917	0.277
	Treatment	Pretest	35	-0.217	0.982	-1.862	1.926
		Posttest	30	-0.051	0.401	-0.884	0.523
23	Control	Pretest	51	-0.509	0.612	-1.563	1.108
		Posttest	54	-0.050	0.250	-0.456	0.533
	Treatment	Pretest	64	0.427	1.222	-1.750	3.043
		Posttest	67	0.319	0.354	-0.497	1.227

Appendix B:
Summary of Efficacy Trial Procedures

Treatment condition (fractions)		Control condition (solving equations)	
Duration	Activity	Duration	Activity
3 hours	• Professional development conducted about a month before gameplay	3 hours	• Professional development conducted about a month before gameplay
30min. × 1 period	• Pretest—administered about 1 week before first day of gameplay • Teacher background survey	30min. × 1 period	• Pretest—administered about 1 week before first day of gameplay • Teacher background survey
40min. × 2 periods	• Game 1: *Wiki Jones* (number line concepts) • Immediate math posttest • Game perception survey • Teacher log	40min. × 2 periods	• Game 1: *Monster Line* (operations on positive and negative integers) • Immediate math posttest • Game perception survey • Teacher log
40min. × 4 periods	• Game 2: *Save Patch* (concepts of unit, fractional pieces, adding fractions) • Immediate posttest • Game perception survey • Teacher log	40min. × 2 periods	• Game 2: *Expresso* (transforming expressions) • Immediate math posttest • Game perception survey • Teacher log
40min. × 2 periods	• Game 3: *Tlaloc's Book* (inverse operations) • Immediate math posttest • Game perception survey • Teacher log	40min. × 4 periods	• Game 3: *Zooples in Space* (solving equations) • Immediate math posttest • Game perception survey • Teacher log
40min. × 2 periods	• Game 4: *Rosie's Rates* (functions, computing slope) • Immediate math posttest • Game perception survey • Teacher log	40min. × 2 periods	• Game 4: *AlgebRock* (solving equations) • Immediate math posttest • Game perception survey • Teacher log
30min. × 1 period	• Posttest—administered about 1 week after last day of gameplay • Teacher survey of experience	30min. × 1 period	• Posttest—administered about 1 week after last day of gameplay • Teacher survey of experience