Using a Single Model Trained across Multiple Experiments to Improve the Detection of Treatment Effects

Thanaporn Patikorn, Douglas Selent, Neil T. Heffernan, Joseph E. Beck, Jian Zou
100 Institute Rd.
Worcester, MA 01609
{tpatikorn, dselent, nth, josephbeck, jzou} @wpi.edu

ABSTRACT

In this work, we describe a new statistical method to improve the detection of treatment effects in interventions. We call our method TAME (Trained Across Multiple Experiments). TAME takes advantage of multiple experiments with similar designs to create a single model. We use this model to predict the outcome of the dependent variable in unseen experiments. We use the predictive accuracy of the model on the conditions of the experiment to determine if the treatment had a statistically significant effect. We validated the effectiveness of our model using a large-scale simulation study, where we showed that our model can detect treatment effects with 10% more statistical power than an ANOVA in certain settings. We also applied our model to real data collected from the ASSISTments online learning platform and showed that the treatment effects detected by our model were comparable to the effects detected by the ANOVA.

Keywords

Intervention Effectiveness; Randomized Controlled Experiments; Meta-Analysis; ANOVA; Treatment Effect; TAME;

1. INTRODUCTION

The goal of this paper is to develop a method that can more effectively detect treatment effects in randomized controlled experiments that are run inside online tutoring systems. Common methods for analyzing these experiments include existing statistical tests such as a T-Test, regression, and an Analysis of Variance (ANOVA). Although these analysis methods are typically used, there are disadvantages that must be considered.

Grossman et al discuss several disadvantages of randomized controlled experiments [4]. One disadvantage is having a small sample size compared to the number of variables and it is unlikely that there will be an equal balance of variables in the control and treatment groups of the experiment. Another disadvantage is that a single study may not be able to infer the overall treatment effect on the entire population. The treatment may have different effects on different subpopulations, experiments settings may be different, and there may also be several different dependent measures to consider. There also may be a large number of experiments where the reported effects are false due to Type I error.

We hope to ameliorate several of these issues by using a technique that combines data from several randomized controlled experiments in order to build a model to estimate the difference between conditions in experiments. Advantages of combining data from multiple experiments include increasing the sample size, and also reducing the variance for better confidence estimates [1].

Two major questions to consider when pooling experiments are discussed in [1]. The first question is, "Which experiments should be combined for analysis?", and is considered "the most serious methodological limitation" [3]. Experiments should be combined if they have similar research questions, populations, experiment settings, intervention components, implementation, and dependent measures. In our paper we select experiments with the same dependent measures and study design format (A/B).

The second question is how to combine experiments once they are chosen for inclusion. One method, called *lumping,* combines all the data into a single data set, ignoring the differences among the experiments. Another method called *pooling*, combines experiments into a single data set but adjusts for differences in experiments [1]. In our case, we have experiments that can have very different effect sizes. We applied the pooling method, but instead of applying standard meta-analysis techniques, we trained a linear model to predict the outcome measures.

Our goal is to use our method called TAME (Trained Across Multiple Experiments) to more effectively detect treatment effects. We use data from multiple experiments to increase the power of the model, and to utilize linear regression to model subject outcomes for treatment effect detection. We hope that TAME would also reduce the bias of meta-analyses in efforts to improve the reliability of statistical results.

The data we use comes from a data set previously collected and synthesized from twenty-two randomized controlled experiments run inside the ASSISTments online tutoring system [5]. These experiments were proposed by internal and external researchers on a large variety of topics. The student population consists of mostly middle-school students ranging from grades 6-8. All experiments had a single control group and a single experiment group (A/B study design) with at least 50 students in each group. A total of 102,252 problems were attempted by 8,297 students across 22 different experiments.

We conducted a large-scale simulation experiment to compare the accuracy of TAME to the accuracy of an ANOVA under different experiment settings. To determine how well each method performed we looked at the chance of detecting an effect when there really is one (true positive) and the chance of not detecting an effect when there really is not one (true negative). This is conversely related to Type I and Type II errors. Our research questions are 1) Does TAME perform better than the ANOVA method? 2) Under what circumstances do TAME perform better?

2. TAME Model

TAME borrows the idea of meta-analysis, where many experiments are used to report on generalized effects. The main concept of TAME is to first model the outcome measure in the absence of the condition assignment. Any other factors can still be used in the creation of the model. To do this, one must use data outside of the experiment of interest (the "test" experiment) to ensure that the model does not overfit to the test experiment. By training a model on a collection of similar experiments, it is less likely that the model will overfit to any given experiment. For the rest of this paper, we will refer to a group of similar experiments as an experiment group.

For each experiment in an experiment group, we first train a linear model on all of the other experiments in the same group, using all factors in the data set except the condition assignments in the experiments. Note that the model used does not have to be a linear model and other types of models will work as well. Once a model is trained, it is applied to estimate the dependent measure of the test experiment. Then, we compute the residual value for each subject in the test experiment, which is the actual outcome measure minus the modeled outcome measure. Assuming that all other factors that may affect the outcome measures are accounted for in the model, the only cause of the residual values must be the condition assignments and noise. A two-tailed unpaired T-test is performed on the residual values of the samples from the control group and the treatment group in the test experiment to determine if there is a significant treatment effect. If the T-test reports that there are significant differences, we claim that the effect of the intervention was statistically significant.

The sign of the residual matters for our usage of the model, which is contrary to most modeling approaches, where the absolute or squared residuals are analyzed. If the residual is positive, it means that the student overperformed the model due to some factors that the model does not account for. Those factors positively affect the student outcome measure and could be attributed to helpful interventions. If the residual is negative, it means that the student underperformed the model, which may be caused by harmful interventions. We believe the reason that our method will result in a better estimate of treatment effects is because training on all experiments except for one, without knowing the conditions of the experiment, will generate a less biased model than an ANOVA, which operates on a single experiment and includes the condition of the experiment while training the model.

3. SIMULATION EXPERIMENT

Simulated data are often used in the EDM community as well as other research areas to validate models , such as [7]. One advantage of using simulated data is that the ground truth values are known, which make it possible to compare the learned values to the true values. Another advantage of using simulated data is that it gives us the ability to control for and test any combinations of parameters. To evaluate the effectiveness of our model, we ran a large scale simulation experiment to compare the accuracy of treatment effects detected by TAME to the accuracy of treatment effects detected by an ANOVA. For both methods, we used a between-subject ANOVA (type III SS) to compare the main effects of the condition variable on our dependent measure using all other factors as fixed factors. We looked at the percent of treatment effects correctly detected (true positive, $p<0.05$) and incorrectly detected (false positive). Our simulation data was generated using Java code and the models were trained and evaluated using R.

Table 1. Parameters, value ranges, and an example of a setting

Parameter	Possible Parameter Values	Example Setting
Expr. in a Group	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	2
Expr. with Diff.	[0, n], n = number of expr. in group	1
Effect sizes	0.05, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1	1.0
Samples	20, 40, 60, 80, 100, 200	20
Factors	0, 1, 2, 3, 4	1
Values per Factor	2, 3, 4	3

3.1. Data Generation

The parameters we experimented with and their possible values are summarized in the first and second column of Table 1, while the third column shows an instantiation of values for an example experiment setting. Ten trials of experimental data were generated for all combinations of parameters resulting in over ten million trials generated.

Experiments in a Group: This parameter represents the number of experiments in a group. We chose to sample groups in the range of [2, 20] experiments in increments of two because we believe this is a realistic number of experiments that could be analyzed together. Several recent meta-analysis papers publish data with the number of studies ranging from 12 - 217 [2, 5, 9]. It is also reasonable to have this many experiments with a similar designs, which can be analyzed together. Our analysis of real data includes a dataset consisting of 22 experiments reported in [5].

Experiments with Differences: This parameter is number of experiments where there is a difference in the outcome measure between the control and treatment group. This value ranges from having no experiments in group with differences to having all the experiments within a group with differences. All experiments that have a difference between the control group and the treatment group all have equal effect sizes.

Samples: This parameter is for the number of samples assigned into a given experiment. In the context of the EDM community, the number samples is equivalent to the number of students that have participated in an experiment. We chose to simulate data for a number of students in the range of {20, 40, 60, 80, 100, and 200} because we believe this range consists of values for a typical number of students expected to participate in most experiments.

Factors: The number of factors for all experiments within an experiment group. The condition of the experiment is considered a special factor and is not grouped with the other factors. All factors are categorical variables. Factors are used to represent features of the student such as gender or levels of prior knowledge, which have been shown to improve predictive modeling [8]. We add features to the generated data to more accurately simulate a real-world scenario. We assume the features do not correlate with the intervention, and therefore do not have interaction effects.

Values per Factor: This parameter represents the number of categorical values that all factors can subsume. For example a factor with two values could represent the gender of a student or a factor with several values could represent the prior knowledge of the student discretized into several bins.

Effect Size: The effect size measured with Cohen's D. Both smaller ranges of differences and larger ranges of differences were tested for both practical and theoretical contexts. In practice many experiments report small effect sizes; therefore we test in the range of [0.05, 0.2] in increments of 0.05 to simulate what would

Table 2. A concrete example of simulated data

Row Number	Experiment Number	Sample Number	Condition	Condition Value	Factor 1	Factor 1 Value	Base Outcome Value	Final Outcome Value
1	1	1	A	0	A	0.4	0	0.4
2	1	2	B	1	B	0.1	0.1	1.2
3	2	2	B	N/A	C	-0.7	0.05	-0.65
4	2	3	A	N/A	A	0.38	-0.7	-0.42

happen in a likely scenario. We also use values from [0.2, 1.0] in increments of 0.2 for larger differences to observe what would happen in a best-case scenario with a large difference in means.

Table 2 shows an example of what the data generated under the example setting in Table 1 looks like. The first column in Table 2 shows what experiment each sample belongs to. In this example there are only two experiments. Each experiment in this example has twenty samples each, however only two samples are shown for both experiments in Table 2. The sample column represents a unique sample number for each experiment. In the context of an experiment, the sample number represents the student. The condition column represents what condition the sample is assigned into. The condition is uniformly and randomly chosen between either "A", or "B", where "A" represents the control group and "B" represents the treatment group. Each condition has a value associated with it, which is equivalent to the effect of the treatment. Table 2 shows that in this example, the intervention has an effect size of 1.0 standard deviation. Therefore the condition value is set to 1.0 where the condition is "B" (treatment), and the condition value is set to 0 where the condition is "A" (control).

Each factor in the experiment has a column for the categorical value of that factor and a value for how that factor value affects the dependent measure of the experiment. Since there is only one factor in this experiment setting, there is only a single factor column ("Factor 1") shown in Table 2. This column can hold three values ("A", "B", or "C"), because the number of values per factor is set to three in this experiment setting. Each factor value is generated randomly and uniformly for each sample. The value for how the factor effects the dependent measure is randomly generated from a standard normal distribution ($\mu = 0$, $\sigma = 1.0$) with Gaussian noise added to the value for each sample for a more realistic simulation. The noise is generated from a normal distribution with the mean centered at the randomly generated value for the factor with a standard deviation of 0.25. In Table 2, this can be seen by looking at rows 1, and 4, which are assigned to factor "A", where all the values for this factor are close to 0.4. In this example the randomly generated effect of factor "A" is 0.4 with noise added for each sample. In the context of educational data mining, certain features of the student can have effects on learning gains which may vary slightly for each student.

The base outcome value is a random number chosen from a normal distribution ($\mu = 0$, $\sigma = 1$). This number represents how a random sample performs. The final column represents the dependent measure in experiments. This value is the sum of the base outcome values, all feature values, and the condition value. For example, row 2 has a condition value of 1, a factor value of 0.1 and a base outcome value of 0.1. Therefore the final outcome value is 1 + 0.1 + 0.1 = 1.2. This representation may be thought of as the average learning gains a student has when comparing their pretest score to their posttest scores. We do not have an explicit dependent measure and will refer to it in the general context.

4. SIMULATION RESULT

To analyze our results we calculated the mean true positive rate and false positive rate at the experiment group level. Each experiment group consisted of a varying number of experiments, with ten trials each. Each trial had a ground truth value where there was either a difference in conditions or there was not a difference in conditions. The ground truth value on whether or not an experiment had differences in conditions is represented in the "experiments with differences" variable described in section 3.1. If a model correctly detected significant differences ($p<0.05$) between conditions it was counted as a true positive. Similarly, if a model incorrectly detected significant differences it was counted as a false positive. An average of the true positive counts and false positive counts for all experiments and trials was used to equally weight each experiment group. Some random data samples generated errors in analysis. If an error occurred for any trial the entire experiment group was removed from analysis to ensure the analysis would be as unbiased as possible. There were 79,200 simulated experiment groups, of which 58 were removed, resulting in 77,842 experiment groups analyzed. The data from the results of the simulation experiment and the code used can be found here. https://sites.google.com/site/tamemethod/

Since there was little change in the false positive rate (Type I error) regardless of method or factors, we exclude it from further analysis. All sets of parameters had a Type I error of roughly 5%, which is the threshold we used to determine if a model detected significant differences. Our analysis focuses on the true positive rates (statistical power) of each method. We ran a repeated measure ANOVA to compare the main effects of the parameters (see data section) on the statistical power of our method to the statistical power of an ANOVA. Out of 70,742 simulated experiments, TAME has an average power of 0.376 (SD = 0.357), which is slightly better than the ANOVA which had an average power of 0.366 (SD = 0.353). This power may seem low, however many experiments in the learning science community do in fact have low power due to the combination of low sample sizes and low effect sizes.

Table 4 shows the results of a repeated measures ANOVA, which determined that the average power of TAME was significantly better than the ANOVA ($F(1, 70{,}713) = 804.144$, $p < 0.001$). We discuss the effect of each parameter in the following sections. We discuss the overall effect each parameter has on both methods and compare the effects between each method.

4.1. Experiments in a Group

There is no general effect of the number of experiments in a group. This is because this variable will only matter for our method which takes advantage of a larger number of experiments in a group when training a model. An ANOVA trains and tests on experiments individually; therefore the number of experiments in group has no effect on the power of the ANOVA. Since the number of experiments has no effect on the power of the ANOVA, it is less likely to see an overall effect considering both TAME and the ANOVA.

Table 3. Tests of Between-Subject Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance	Partial Eta Squared
Intercept	3285.103	1	3285.103	115891.122	< 0.001	0.621
effect size	13753.480	7	1964.783	69313.168	< 0.001	0.873
factors	37.834	4	9.459	333.678	< 0.001	0.019
values per factor	1.196	2	0.598	21.096	< 0.001	0.001
samples	2013.213	5	402.643	14204.334	< 0.001	0.501
experiments	0.163	9	0.018	0.638	0.765	0
percent of exp. with diff.	0	1	0	0.011	0.917	0
Error	2004.463	70713	0.028			

Table 4. Test of Within-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance	Partial Eta Squared
method	0.576	1	0.576	804.144	< 0.001	0.011
method * effect size	2.948	7	0.421	587.633	< 0.001	0.055
method * factors	2.205	4	0.551	769.046	< 0.001	0.042
method * values per factor	0.671	2	0.335	467.906	< 0.001	0.013
method * samples	1.173	5	0.235	327.340	< 0.001	0.023
method * experiments	0.050	9	0.006	7.709	< 0.001	0.001
method * percent of exp. with diff.	0.002	1	0.002	2.468	0.116	0
error(method)	50.683	70713	0.001			

There is also no overall noticeable difference between TAME and an ANOVA for different number of experiments in a group. Table 3 shows that the number of experiments in a group has a significant effect on power ($F(9,70713) = 7.71$, $p<0.001$) with a partial eta squared = 0.001. Although the difference between the two methods is statistically significant, the effect size is insignificant.

Although there is no overall difference in method type for varying the number of experiments in a group, the number of experiments has a major impact in the case where there are a large number of factors and a small number of samples with a high effect size. Figure 1 shows that for a subset of experiments, as the number of experiments in a group increases, the difference in power between the two methods increases. TAME has a power of 0.27 compared to a power of 0.22 for the ANOVA with two experiments in a group and TAME has a power of 0.35 compared to a power of 0.25 for the ANOVA with ten experiments in a group.

4.2. Number of Factors

More factors introduce more noise in the data, making it harder to detect treatment effects. Table 3 shows that the number of factors has a significant effect on power ($F(4,70713) = 333.67$, $p<0.001$) with a partial eta squared = .019. Figure 2 shows that as the number of factors increases, the power of TAME decreases less than the power of ANOVA. This decrease leads to a difference in power between the two methods based on the number of factors. The number of factors is statistically significant ($F(4,70713) = 769.046$, $p<0.001$) with a partial eta squared of 0.042. We believe this is because TAME accounts for noises better than ANOVA by using more data that is available to TAME.

4.3. Number of Samples

In general, more samples lead to a better estimate of the true means and more power. Table 3 shows that the number of samples has a significant effect on power ($F(5,70713) = 14204.334$, $p<0.001$) with a partial eta squared = 0.5. As the number of samples increases, both methods perform equally well. This result is expected.

Table 4 shows that TAME performs better slightly than the ANOVA when there are a fewer number of samples, since the ANOVA is not an optimal method in this situation. The number of samples is a statistically significant factor when comparing the power differences between the two methods ($F(5,70713) = 327.340$, $p<0.001$) with a partial eta squared of 0.023.

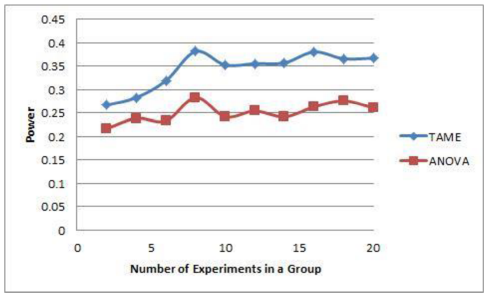

Figure 1. The power as the number of experiments in a group increases for experiment groups with 20 samples, four factors, and a treatment effect size of 0.8 and 1.0.

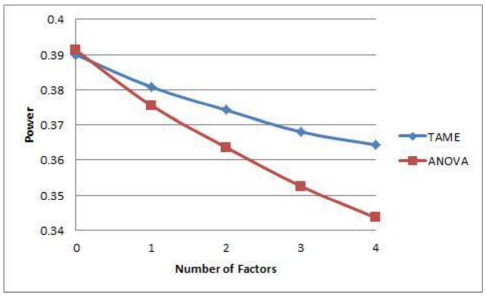

Figure 2. The statistical power of TAME and ANOVA by the number of factors used to train the models

4.4. Effect Size

A larger treatment effect is easier to detect and therefore has a positive impact on power. Table 3 shows that the size of the treatment effect has a significant effect on power ($F(7,70713) = 69313.168$, $p<0.001$) with a partial eta squared = 0.873. As the size of the effect increases so does the power.

Table 4 shows TAME performs slightly better than the regular ANOVA as the treatment effect increases. The effect size is a statistically significant factor when comparing power differences between TAME and ANOVA, ($F(7,70713) = 587.633$, $p<0.001$) with a partial eta squared of 0.055.

5. REAL DATA RESULT

We applied both TAME and the ANOVA method on a data set composed of twenty-two randomized controlled experiments run inside the ASSISTments online learning platform to compare the two method on real data [6]. Every experiment in the group is a Skill Builder consisting of one control group and one treatment group. A Skill Builder is "an assignment type that consists of a large number of similar problems, where students must answer a specified number of problems (usually three) correctly in a row on the same day in order to finish the assignment." [6]. We applied both TAME and an ANOVA on students in the studies, with the following factors as training factors: Prior Percent Correct, Guessed Gender, Prior Percent Completion, Z Scored Mastery Speed, Prior Homework Percent Completion, Z Scored HW Mastery Speed. For dependent measure, we use logarithm with base ten of the Mastery Speed, which is the number of problems a student took to answer three problems correctly in a row [9]. We use the logarithm of Mastery Speed to reduce the effect of outliers.

Table 6 shows that our method can be applied to detect significant different between conditions of a real data set. Since the size of each experiment in the data set is greater than 100, the result of simulation study suggests that TAME is as good at detecting significant differences as ANOVA. Both TAME and ANOVA detected significant differences between conditions of the same experiments (2, 3, 4, 10, and 22). This result further supports our claim that TAME is a good alternative to ANOVA, if not better.

We further investigated the reliability of TAME and ANOVA. For each experiment, we trained a model using all of the data from the other twenty-one experiments. We then used this model to predict the performance on the data in the test experiment. We experimented with a different sample size of (10, 20, 30, 40, 50, 60, 70, 80, 90, and 100) to predict in the test experiment. The evaluation of each model was an average of running the model 1,000 times, with a different random set of data points in the test experiment each time. This methodology does not invalidate our analysis since TAME was designed to utilize all data from outside of the target experiment, such as data from experiments in the past, and such data are not affected by the sample size of the target experiment. We chose to report on the results of two of the experiments in Table 5 and Table 7.; experiment 3, which was the experiment that we found the strongest treatment effect for, and experiment 6, which was one of the experiments that we did not find a significant treatment.

Table 5. The probability and the confident interval of detecting the treatment effect on the resampled data set ($p < 0.05$) on experiment 3

Experiment 3	Probability of Detecting Treatment Effect ($p<0.05$)		Size of Adjusted Wald Confidence Interval	
sample size	TAME	ANOVA	TAME	ANOVA
10	0.2280	0.0650	0.0260	0.0154
20	0.4410	0.3360	0.0307	0.0292
30	0.5810	0.5590	0.0305	0.0307
40	0.7070	0.7050	0.0282	0.0282
50	0.7970	0.8020	0.0249	0.0247
60	0.8580	0.8530	0.0217	0.0220
70	0.9170	0.9180	0.0172	0.0171
80	0.9420	0.9580	0.0147	0.0127
90	0.9660	0.9610	0.0115	0.0122
100	0.9710	0.9810	0.0107	0.0088

Table 6. Summary statistics and significance for the real dataset

	Mastery Speed Control and Experiment Group	Mastery Speed Control Group	Mastery Speed Experiment Group	TAME Sig.	ANOVA Sig.	ANOVA Partial Eta Squared
1	$\mu = 0.80, n = 468, \sigma = 0.21$	$\mu = 0.79, n = 256, \sigma = 0.20$	$\mu = 0.82, n = 212, \sigma = 0.22$	0.208	0.222	0.003
2	$\mu = 0.78, n = 672, \sigma = 0.24$	$\mu = 0.76, n = 324, \sigma = 0.21$	$\mu = 0.80, n = 348, \sigma = 0.26$	0.014	0.013	0.009
3	$\mu = 1.16, n = 240, \sigma = 0.12$	$\mu = 1.12, n = 123, \sigma = 0.11$	$\mu = 1.21, n = 117, \sigma = 0.11$	0.000	0.000	0.162
4	$\mu = 1.12, n = 540, \sigma = 0.21$	$\mu = 1.10, n = 298, \sigma = 0.19$	$\mu = 1.16, n = 242, \sigma = 0.22$	0.001	0.001	0.020
5	$\mu = 0.67, n = 1303, \sigma = 0.25$	$\mu = 0.67, n = 667, \sigma = 0.25$	$\mu = 0.67, n = 636, \sigma = 0.24$	0.503	0.389	0.001
6	$\mu = 0.63, n = 337, \sigma = 0.17$	$\mu = 0.62, n = 165, \sigma = 0.18$	$\mu = 0.63, n = 172, \sigma = 0.16$	0.634	0.737	0.000
7	$\mu = 0.65, n = 365, \sigma = 0.16$	$\mu = 0.65, n = 202, \sigma = 0.16$	$\mu = 0.65, n = 163, \sigma = 0.16$	0.489	0.562	0.001
8	$\mu = 0.59, n = 455, \sigma = 0.17$	$\mu = 0.59, n = 223, \sigma = 0.18$	$\mu = 0.59, n = 232, \sigma = 0.16$	0.542	0.571	0.001
9	$\mu = 0.91, n = 119, \sigma = 0.16$	$\mu = 0.93, n = 52, \sigma = 0.18$	$\mu = 0.90, n = 67, \sigma = 0.14$	0.460	0.478	0.005
10	$\mu = 1.09, n = 432, \sigma = 0.20$	$\mu = 1.07, n = 212, \sigma = 0.18$	$\mu = 1.11, n = 220, \sigma = 0.22$	0.037	0.045	0.010
11	$\mu = 0.95, n = 171, \sigma = 0.20$	$\mu = 0.96, n = 84, \sigma = 0.21$	$\mu = 0.93, n = 87, \sigma = 0.19$	0.297	0.225	0.009
12	$\mu = 0.90, n = 122, \sigma = 0.18$	$\mu = 0.92, n = 60, \sigma = 0.19$	$\mu = 0.88, n = 62, \sigma = 0.16$	0.302	0.389	0.007
13	$\mu = 1.11, n = 148, \sigma = 0.24$	$\mu = 1.08, n = 70, \sigma = 0.28$	$\mu = 1.12, n = 78, \sigma = 0.21$	0.320	0.395	0.005
14	$\mu = 0.83, n = 174, \sigma = 0.16$	$\mu = 0.84, n = 99, \sigma = 0.17$	$\mu = 0.82, n = 75, \sigma = 0.14$	0.159	0.216	0.009
15	$\mu = 0.93, n = 240, \sigma = 0.19$	$\mu = 0.94, n = 124, \sigma = 0.19$	$\mu = 0.91, n = 116, \sigma = 0.19$	0.159	0.177	0.008
16	$\mu = 0.98, n = 121, \sigma = 0.17$	$\mu = 0.97, n = 63, \sigma = 0.14$	$\mu = 1.00, n = 58, \sigma = 0.19$	0.159	0.324	0.009
17	$\mu = 0.94, n = 226, \sigma = 0.18$	$\mu = 0.95, n = 120, \sigma = 0.17$	$\mu = 0.94, n = 106, \sigma = 0.20$	0.529	0.342	0.004
18	$\mu = 0.70, n = 264, \sigma = 0.13$	$\mu = 0.71, n = 126, \sigma = 0.14$	$\mu = 0.70, n = 138, \sigma = 0.13$	0.455	0.226	0.006
19	$\mu = 1.02, n = 218, \sigma = 0.19$	$\mu = 1.02, n = 105, \sigma = 0.17$	$\mu = 1.02, n = 113, \sigma = 0.20$	0.844	0.994	0.000
20	$\mu = 0.81, n = 825, \sigma = 0.18$	$\mu = 0.81, n = 409, \sigma = 0.19$	$\mu = 0.81, n = 416, \sigma = 0.17$	0.887	0.926	0.000
21	$\mu = 0.84, n = 291, \sigma = 0.15$	$\mu = 0.85, n = 140, \sigma = 0.15$	$\mu = 0.84, n = 151, \sigma = 0.15$	0.855	0.892	0.000
22	$\mu = 0.78, n = 213, \sigma = 0.16$	$\mu = 0.80, n = 111, \sigma = 0.16$	$\mu = 0.75, n = 102, \sigma = 0.15$	0.020	0.018	0.027

Table 7. The probability and the confident interval of detecting the treatment effect on the resampled data set (p < 0.05) on experiment 6

experiment 6	probability of detecting treatment effect (p<0.05)		Size of Adjusted Wald Confidence Interval	
sample size	TAME	ANOVA	TAME	ANOVA
10	0.0560	0.0340	0.0144	0.0115
20	0.0500	0.0450	0.0137	0.0131
30	0.0510	0.0660	0.0138	0.0155
40	0.0470	0.0650	0.0133	0.0154
50	0.0560	0.0580	0.0144	0.0147
60	0.0720	0.0780	0.0162	0.0167
70	0.0540	0.0540	0.0142	0.0142
80	0.0490	0.0530	0.0136	0.0141
90	0.0620	0.0670	0.0151	0.0156
100	0.0520	0.0600	0.0139	0.0149

Table 5 shows that for the experiment with the strongest treatment effect (experiment 3), TAME is able to detect the treatment effect better than ANOVA, especially when the sample size <= 40. This result agrees with the result of our simulation study. When the treatment effect is not present (experiment 6), the false positive rate of both TAME and ANOVA are around 5% as shown in Table 7. This result is to be expected from using a p-value threshold of 0.05.

6. CONTRIBUTIONS

This paper makes three contributions. The first contribution of this paper is TAME, a more robust and more effective method of detecting treatment effects that can analyze several experiments simultaneously. Since the TAME model is not built specifically for any particular experiment, it allows the same model to generalize to experiments unseen by the model, including future experiments. To our knowledge, this is the first method that detects treatment effects on multiple experiments individually and simultaneously.

The second contribution this paper makes is that the results from a large-scale simulation experiment showed that TAME is better at detecting treatment effects compared to an ANOVA by more than ten percent in the case where there is a large effect, fewer samples, more factors, and with more experiments. This simulation experiment validated our proposed method and also showed that TAME has slightly better statistical power than an ANOVA and never performs worse. TAME can quickly detect large differences, such as when the treatment is harmful. It is important to detect harmful interventions as soon as possible to ensure that students are exposed to the least amount of negative effects.

The third contribution this paper makes is taking our validated method and applying it to real data collected from twenty-two randomized controlled experiments run in the ASSISTments online learning platform. On this data set, TAME and ANOVA are in agreement on significant differences between conditions. This result allows the associated researchers to further investigate the interventions and their effects, allowing them to better understand how students learn and, eventually, develop better tools and interventions for students.

6.1. Future Work and Conclusions

This work is a first step in building a model that can be used across interventions to estimate effect sizes. As such, there are many future directions to explore. A possible future work involves equally weighting the experiments our model uses. It is rare for all experiments to all have the same number of samples. Currently our model gives more weight experiments with more samples. This may lead to a small number of experiments accounting for a large amount of the weight when training a model. For future work the weighting of experiments and the effect can be investigated.

In conclusion, we have created a single model that generalizes across experiments. We have shown how it can be applied to multiple, unseen, experiments in order to evaluate their efficacy. This approach is in contrast to creating separate models for each intervention we are evaluating. This model is able to detect the effect of each intervention relative to other interventions and provide a set of features that might affect and interact with interventions. In addition, the same trained model can be applied to investigate future interventions. We evaluated the effectiveness of our model in a simulation study, which shows that our model can detect significant differences 10% more than an ANOVA in certain cases. We then applied our model to real data and found that three out of twenty-two interventions are significantly different from the control conditions.

7. ACKNOWLEDGEMENTS

We thank multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DGE-1535428 & DRL-1031398), the US Dept. of Ed (IES R305A120125 & R305C100024 and GAANN), and the ONR.

8. REFERENCES

[1] Bangdiwala, S. I., Bhargava, A., O'Connor, D. P., Robinson, T. N., Michie, S., Murray, D. M., & Pratt, C. A. (2016). Statistical methodologies to pool across multiple intervention studies. *Translational Behavioral Medicine*, 1-8.

[2] D'Mello, S.K. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology 105*, 4, 1082–1099

[3] DeMets, D. L. (1987). Methods for combining randomized clinical trials: strengths and limitations. *Statistics in medicine*, *6*(3), 341-348.

[4] Grossman, J., & Mackenzie, F. J. (2005). The randomized controlled trial: gold standard, or merely standard?. *Perspectives in biology and medicine*, *48*(4), 516-534.

[5] Patall, E.A., Cooper, H., & Robinson, J.C. (2008). The Effects of Choice on Intrinsic Motivation and Related Outcomes: A Meta-Analysis of Research Findings. *Psychology Bulletin. 134* (2), pp 270-300.

[6] Selent, D., Patikorn, T., & Heffernan, N. (2016). ASSISTments Dataset from Multiple Randomized Controlled Experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 181-184). ACM.

[7] Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. Journal of Educational Measurement, 47(2), 150-174.

[8] Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. Journal of Educational Psychology,105(4), 932.

[9] Xiong, X., Li, S., & Beck, J. E. (2013). Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle. In FLAIRS Conference.