120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #7812

Measuring the effectiveness of pedagogical innovations using multiple baseline testing

Mr. Alex Albert, University of Colorado

Alex Albert is a PhD Candidate in the Construction Engineering and Management Program at the University of Colorado at Boulder. He has conducted research for the Construction Industry Institute and ELECTRI International, studying hazard recognition and response. Alex specializes in implementing experimental research methods in engineering education to perform hypothesis testing and draw causal inferences.

Dr. Matthew R. Hallowell, University of Colorado

Dr. Matthew Hallowell is an Assistant Professor at the University of Colorado at Boulder. He earned a PhD with a dual focus in Construction Engineering and Management and Occupational Safety and Health from Oregon State University. He also earned a BS and MS in Civil Engineering from Bucknell University. For his efforts in teaching innovations, Dr. Hallowell has received the National Science Foundation CAREER award, Beavers Endowed Faculty Fellowship, John and Mercedes Peebles Innovation in Education Award, the ASCE New Faculty Excellence in Teaching Award, the Department of Civil, Architectural, and Environmental Engineering's Teaching Award, University of Colorado College of Engineering and Applied Science Dean's Performance Award, and the Best Technical Publication Award from the Journal of Safety, Health, and Environmental Research for a paper on a new teaching innovation. He has also selected as an ASCE ExCEEd Teaching Fellow Teaching Mentor and as a President's Teaching and Learning Collaborative Member.

Measuring the effectiveness of pedagogical innovations using multiple baseline testing

ABSTRACT

A great deal of literature focuses on innovations that are designed to improve educational performance. Although some innovations are designed and implemented to address learning in a very specific domain, others influence student learning more generally as they are applicable regardless of specific content (e.g., mechanisms for delivering new content, new strategies for student-student interactions, and application of new technologies). Many instructors form the hypothesis that a particular innovation will enhance student learning and, consequently, the ability to achieve desired learning objectives. Testing such a hypothesis can be troublesome when confounding factors exist in the student body's learning environment such as scheduled breaks, social stressors, and activities occurring in other courses. Multiple baseline testing is a promising strategy for statistically controlling the influence of confounding factors when innovations are implemented consistently across multiple groups of students. This strategy involves measuring student performance, implementing the innovation at a randomly selected time, and continuing to measure student performance as the innovation is integrated within the course. The impact of the innovation treatment can be measured using time series regression. This paper presents the proper mechanics of multiple baseline testing, discusses the relatively small body of research on this method that exists outside the medical and biological fields, and provides clear recommendations for managing threats to validity in engineering education research.

Introduction

In much of pedagogy literature authors attempt to describe a pedagogical innovation and demonstrate its impact on student learning. These studies include qualitative measurements of improvement such as student feedback in learning logs [1] and quantitative measurements such as performance on examinations [2]. The vast majority of researchers assess the impacts of new teaching methods primarily using correlational or comparative studies. They often gather empirical data to understand ***if*** there is an improvement combined with qualitative feedback in student reflections to understand ***why*** the intervention was successful or unsuccessful. Nearly all of these pedagogical studies aim to measure the improvement in learning resulting from an intervention. These studies essentially aim to perform a hypothesis test (i.e., testing to see if the implementation of intervention X yields a statistically significant improvement in achievement of learning objectives) to infer causal relationships. The problem with such causality inference approaches is that these methods can inherently be susceptible to limitations in internal and external validity as there are numerous confounding factors that may influence achievement of learning objectives including instructor effectiveness, social stress, time of the year, and others.

Although several correlational studies have claimed to indicate causal relationships in education research [3,4], several researchers rightly question the legitimacy of such studies [5-8]. According to these researchers, a causal inference can only be inferred if the following criteria are warranted:

- Sufficient evidence that the effect or outcome variable occurs as a consequence of introducing a specific treatment variable;

- Clear indication of the absence of any alternate plausible explanation for the effect observed; and
- Evidence that the causal factor or treatment variable precedes the occurrence of the observed effect

In light of the above requirements, researchers posit that correlation based cross-sectional research that measures outcome variables at a single point in time inherently fails to provide adequate evidence for causal inference. In fact, it is impossible to provide evidence to assert that the causal factor preceded the occurrence of the observed effect [8]. Also, such studies do not adequately control for extraneous or alternate plausible explanations for the observed effect [9]. Ironically, results from a 2004 survey of five teaching and learning journals by Robinson et al. (2007) indicate that 43% of non-intervention studies contained causal statements[6]. Such trends have led Hsieh et al. (2005)[5], Seethaler and Fuchs (2005)[10], and Robinson et al. (2007)[6] to express concern with research rigor. They encourage education researchers to reinvigorate their intervention research undertakings. Fortunately, there are experimental and quasi-experimental methods that can achieve validity and should be used to make valid causal inferences. As noted by Thompson et al. (2005)[11], randomized controlled intervention experiments are a requirement for providing definitive answers to causal questions.

Randomized controlled intervention studies are true experiments in which subjects are randomly assigned to at least two conditions, namely the intervention or treatment group and a control group. The researcher intentionally manipulates or introduces the treatment variable to the intervention group [12]. The control group, which does not receive the intervention, is compared to the treatment group to compute effects of the independent variable. Accordingly, causal inferences based on the difference in the observed outcome between the treatment and the control group can be attributed to the intervention. As such, randomized controlled intervention studies systematically account for or eliminate alternate plausible explanations enabling definitive casual inferences [13].

In educational research, contrary to intuition, it has been established that the number of articles based on randomized experiments have considerable declined over the years. According to Hsieh et al. (2005) [5], the results from surveying 4 educational journals indicate that the percentage of educational articles featuring randomized experiments decreased from 47% in 1983 to 34% in 1995, and to only 26% in 2004. In another study conducted by Snyder et al. (2002) [14], in a review of 450 group quantitative studies, only 10% represented randomized controlled experiments. This decline in randomized experiment studies may partly be attributed to the following factors: (1) randomized designs rarely duplicate real-life situations [15]; (2) practical conditions for randomized experiments are generally not satisfied [16]; (3) the randomization process may be especially challenging in an educational setting where study groups may not be altered to form comparable intervention and control groups; and (4) ethical considerations emerge when a promising or potential educational intervention is provided to the intervention group while the control group is denied of its benefits [17]. Interestingly, the decline in proportion of experimental education studies has occurred despite the fact that several legislations (e.g. No Child Left Behind - NCLB 2001) and authors have elevated randomized experiments as being the “gold standard” for conducting scientifically credible research [6,13,18].

One major reason for the decline in the number of intervention studies is the perception among researchers that the required methodological rigor to conduct scientifically credible conclusions is impractical in an educational setting [5,19,20]. As discussed, however, correlation-based studies have been criticized because definitive causal inferences cannot be established. Therefore, there is an imminent need in the field of educational research to understand how to conduct rigorous research that yields valid causal interferences that is also practical.

A method with great potential in the pedagogical domain for experimental research is multiple baseline testing (MBT). This experimental technique allows a researcher to conduct a controlled and internally valid experiment when a longitudinal assessment strategy is practical. Although MBT is time intensive, the method is rigorous because its inherent structure limits threats to validity and reliability and allows the researcher to make valid causal inferences [21]. This highly potential research design remains underused despites its ability to produce scientifically reinforced results in educational research [22]. The objective of this paper is to describe the MBT method, how to form hypotheses that are appropriate for MBT, how to structure a proper MBT experiment, methods for promoting validity and reliability during the MBT process, proper statistical approaches for time series data subject to autocorrelation. We present this guidance in the context of six experiments conducted in professional research and two experiments conducted in the classroom. We expect that the guidance provided can be used by future investigators to increase the rigor of their pedagogical research and to serve as a foundation for experimental research for establishing causal relationships. At the present time, there is no singular resource for the proper use of MBT for educational research despite its utility, practicality, and rigor for drawing causal inferences regarding improvements resulting from pedagogical innovations. Thus, this paper should be of interest for researchers across all pedagogical domains.

Background and Rationale of MBT

The MBT design methodology was first introduced in the *Journal of Applied Behavior and Analysis* by Baer et al. (1968) [23] where the authors argued that the effects of experimental manipulations, if any, could be definitively illustrated with the MBT structure (as cited in [24,25]. Since then, several research methodologists have recommended the use of the MBT design to evaluate the effectiveness of interventions in various fields [26-29]. The medical and biological fields in particular have realized the benefit of MBT [30,31]. Unfortunately, it remains underused in education research [22].

Several methods exist to test the effects of pedagogical interventions. Since the implementation of true randomized experiments that deprive the control group of potential interventions is often considered unethical [13], researchers are forced to use quasi-experimental techniques involving repeated measures that provide each participant the benefits of the intervention. As a result, methods that include pure control groups such as 'Pretest–posttest randomized or non-randomized experimental control group' designs are often disregarded while conducting pedagogical intervention experiments [32,33]. Further, the use of a simple pre-test and post-test design methods do not portray change over time because data are gathered at only two instances in time (before and after). These designs, known as AB designs, provide very poor internal validity as unrelated confounding variable that occurred during the intervention could potentially distort results [34]

Currently, there is a visible increase in the use of single or group time-series designs. In such designs, repeated longitudinal data are gathered over time. This method permits researchers to make reliable causal inferences based on the baseline logic [35]. According to this logic, repeated measures taken under at least two different conditions: baseline and intervention phase could be compared to measure effect changes. In fact, according to several authors, longitudinal data are the most reliable and rigorous approach to measure change [6,36,37]. Moreover such methods are attributed to possess several advantages such as increased internal validity as inter-participant or group differences do not systematically distorted inferences and repeated measures reduce measurement errors [38].

The MBT design, illustrated in Figure 1, involves multiple A-B (before and after) design studies that are conducted simultaneously to enhance validity and reliability of inferences [39,40]. Specifically, repeated baseline measurements are simultaneously gathered across independent groups of students to represent the performance prior to the introduction of the intervention. Following this, interventions are introduced to each student group on a staggered basis [41-43]. That is, after gathering adequate baseline measurements for one student group, the intervention is introduced to the group while the other group(s) are maintained at their baselines. This process is repeated until all groups are introduced to the interventions. As such, all students participating in the study receive the potential intervention, thus avoiding any ethical considerations [44].

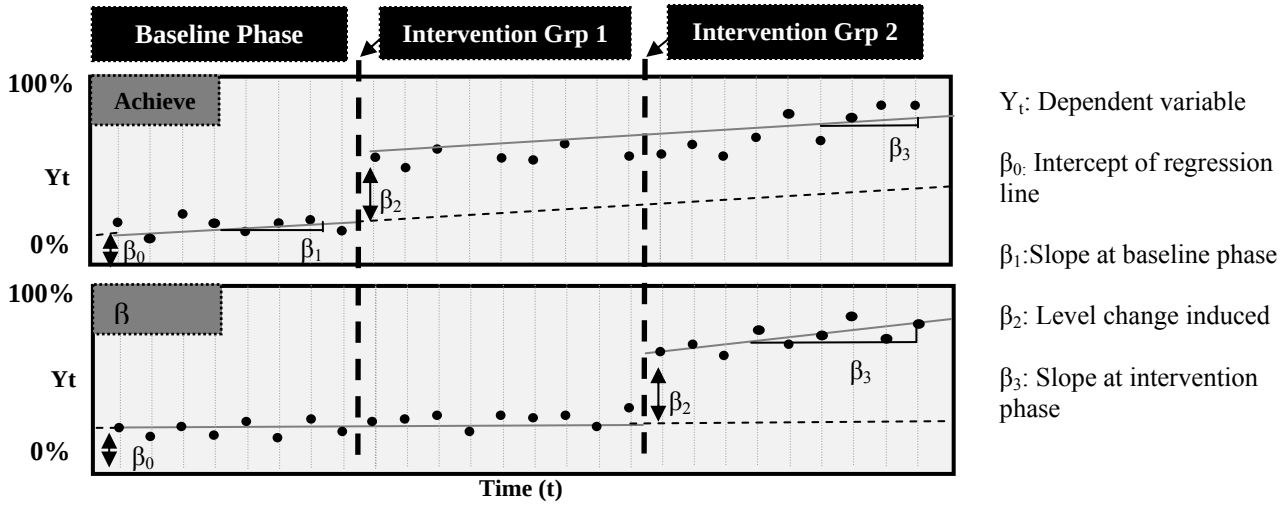


Figure 1: Schematic representation of the MBT design (Y_t: proportion of hazards recognized)

The MBT design methodology not only allows comparison of the dependent variable within baselines, but also allows comparisons among baselines. Therefore, while one group is introduced to the treatment, the other groups perform as controls. Also, the measurements at the baseline phase for each group provides an additional control during the intervention phase [29]. As such, the simultaneous baseline measurements allow the researcher to verify that changes in the dependent variable occurs only as a result of introducing the intervention or treatment variable. Thus, inferences are reinforced by the observed changes in the dependent variable across multiple baselines [27]. Finally, exchangeability or equivalence of the groups in the study is not a concern as comparisons of effect change is made against the performance of the group in the

baseline phase. Such a design also excludes between subjects sources of variability, thus providing better estimates of effect size [45].

Implementation of MBT in Applied Settings

Practical methods of experimental research such as the single or group time-series designs have been published for years in reputed medical and behavioral journals [46]. However, true experiments that draw causal inferences are less common in education research [47]. This knowledge gap leads to the frequent and inappropriate use visual analysis in lieu of statistics, which do not yield scientifically credible results [29,48,49]. In fact, Huitema and Mckean (2000) claim that most inferential techniques currently used are egregious. MBT offers a potential solution for the applied research domain because the technique provides valid inferences while being intuitive. The challenge with MBT is determining if MBT is the correct method for a specific hypothesis or research question, proper design, and statistical techniques.

Analysis Method Considerations

The most simple and intuitive analytical method for MBT is to consider each baseline as independent two- phase AB studies and measure effect sizes across student groups. Therefore, for each baseline, data in the pre-intervention phase provides information regarding the level of performance and associated variability prior to intervention introduction. Similarly, in the post-intervention phase, the level of performance and its associated variability may be observed, which resulted due to the intervention. In addition to the change, level slope trend and the existence of serial dependency may also be observed from both phases.

Once the longitudinal data are collected the researcher may test several hypotheses to determine if there is sufficient statistical evidence to infer that a change in performance has occurred as a result of the intervention. In order to test such hypotheses, the researcher must consider two main effect sizes: level change and slope change [50]. Both of these estimates for intervention analysis can be computed by several methods but literature predominantly uses *ARIMA Intervention Models* or *Time-Series Regression Models.* Since *ARIMA models* require a minimum of 50 to 100 observations to make valid inferences, it may be impractical in small classroom settings. Thus, the remainder of the paper focuses *Time-Series Regression Models* because they apply to both large and small samples.

Effect size determination

Level change (β_1) as shown in Figure 1, refers to immediate changes in the dependent variable once the intervention has been introduced. It is the difference between the dependent variable computed by extrapolating data form observations in phase 1 and the expected value of the dependent variable computed from the phase 2 dataset at a specific time, *t*. In other words, level change is the difference between the predicted value of the dependent variable based on the pre-intervention regression line and the post intervention line at time, *t*. It is important to note that the comparisons are made based on the same point in time, which usually corresponds to the first observation after intervention introduction.

Slope change (β_3), on the other hand, refers to delayed effects. That is, slope change the difference in slope between the best fit baseline and post-intervention line, which is estimated by linear regression. These two effect sizes are adequate to describe changes in performance

between the pre-intervention and post-intervention phase. It is common to have interventions that depict both level change (β_1) and slope change ($\beta 3$) as depicted in Figure 1.

Development Research and Statistical Hypothesis

Every inferential study begins with developing the research and the statistical hypothesis. These are often assumptions that need to be tested. Researchers that are interested in testing the effects of an intervention, typically design their intervention based on knowledge and established literature to provide strong theoretical evidence that the hypothesis may be true. The statistical hypothesis consists of the null hypothesis that states the negation of what the researcher expects to observe, while the alternate hypothesis states that the predicted results are probabilistically true. For example, a study involving a pedagogy intervention would be designed to test a null hypothesis that: *the introduction of the intervention will not result in higher degree of retention than the traditional method.* Statistical analytic methods are designed to establish the probability of the null hypothesis being true. If the probability of the null hypothesis to be true is shown to be improbable (based on acceptable probability, α), then the alternate hypothesis is accepted.

For testing the effects of a specific intervention, based on the above listed effect sizes, the two statistical hypotheses that will essentially be of interest in an MBT study are shown as Equation 1 and 2.

$$\textbf{Null hypothesis: } \beta 1 = 0; \textbf{ Alternate hypothesis: } \beta 1 \neq 0 \qquad (1)$$

$$\textbf{Null hypothesis: } \beta 1 = \beta 3 = 0 = 0; \textbf{ Alternate hypothesis: } \beta 1 = \beta 3 \neq 0 \qquad (2)$$

As indicated in Equation 1, the null hypothesis assumes that there is no level change indicating the absence of an immediate change after the intervention is introduced. Similarly, in Equation 2, the null hypothesis assumes that there is no slope in either the pre-intervention phase or the post-intervention phase. If a slope does exist in either phase, the null hypothesis is rejected and the alternate hypothesis is accepted. Then, based on the mathematical model used for analysis the estimates for β_1 and β_3 are computed and the sum of the two would yield the slope in the post-intervention phase.

Models for analyzing Time-Series Regression Intervention Analysis

As described above, the analysis procedure will treat each baseline as an independent two-phase AB design and the overall effect size for the MBT design will be determined by integrating results from individual baselines. The first step in the analysis involves the selection of an appropriate statistical model to represent the observed data. Of several mathematical models in literature the model suggested by Huitema and McKean [43,51] shown in Equation 3 is specifically appropriate for the determination of the effect sizes. In cases where the null hypothesis ($\beta 1 = \beta 3 = 0$) is satisfied, the equation can be reduced to Equation 4. Hence, if $\beta 1 = \beta 3 = 0$, then modeling the data based on Equation 4 would yield higher power for analysis purposes.

If the data do not meet the assumption of independence (i.e., they are autocorrelated), then a modified relationship must be considered as shown in Equation 5 [52]. Observations are said to be autocorrelated if errors at a given time is associated or can be used to predict errors at a future time. Similar to the above case, if the null hypothesis $\beta 1 = \beta 3 = 0$ is satisfied, then the equation can be reduced to as shown in Equation 6.

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \varepsilon_t \quad (3)$$

$$Y_t = \beta_0 + \beta_2 D_t + \varepsilon_t \quad (4)$$

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \varphi_1 \varepsilon_{t-1} + u_t \quad (5)$$

$$Y_t = \beta_0 + \beta_2 D_t + \varphi_1 \varepsilon_{t-1} + u_t \quad (6)$$

Where, Y_t is the dependent variable at time t; β_0 is the intercept of the regression line at t = 0; β_1 is the slope at the baseline phase; β_2 is the level change measured at time n_1+1; β_3 is the change in slope from the baseline phase to the intervention phase; T_t is the value of the time variable T at time t; Dt is the value of the level-change dummy variable D (0 for the baseline phase and 1 for the intervention phase) at time t; SC_t is the value of the slope-change variable SC defined as [Tt -(n1 + 1)]D; n1 is the number of observations in the baseline phase; ε_t is the error of the process at time t; φ_1 is the lag-1 autoregressive coefficient; u_t is $Y_t - (\beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \varphi_1 \varepsilon_{t-1})$ at time t.

Analysis Methodology

Step 1: Test of Assumptions for Regression Analysis

Before beginning regression analysis as with most statistical procedures, the underlying assumptions need to be verified. Violations of these assumptions can yield biased or spurious results. Specifically, assumptions of homoscedasticity of residuals, normality of errors and, independence of errors (autocorrelation) is to be tested. Methods to test and remedy violations for each assumption can be found in any elementary regression analysis textbook, and hence is not addressed here. However, if the assumption of independence of errors is violated, then Equation 5 or 6 can be utilized to account for autocorrelation. Tests such as the Durbin-Watson test statistic and the Huitema-McKean test of autocorrelation can be effectively used to test the lag-1 autocorelation among the observations.

Step 2: Selection of Mathematical model

As mentioned above based on the test for autocorrelation, the appropriate models is to be chosen. That is, if the test for autocorrelation ($\rho=0$) is accepted, Equations 3 and 4 must be compared to choose the appropriate model .On the other hand if $\rho \neq 0$, Equation 5 and 6 are to be compared. For example, if $\rho=0$ is accepted, the parameters of the regression Equation 3 and 4 are computed by regressing the dependent variable (Y) on the respective predictor variables. Once the parameters of the two equations (3 and 4) are computed the hypothesis, $\beta 1 = \beta 3 = 0$ is tested using the test statistic shown in Equation 7.

$$F = \frac{SS_{Reg\ ModelI} - SS_{Reg\ ModelII}}{MS_{Reg\ ModelI}} \quad (7)$$

If the hypothesis $\beta 1 = \beta 3 = 0$ is rejected, then Equation 3 is selected as being the most appropriate model for the data; else Equation 4 is selected. The same procedure is followed for the autocorrelated models. Finally, the regression coefficients of the selected model represent level change (β_1) and (β_3), if any.

Step 3: Computation of the Overall Effect size for the MBT study
Based on the effect sizes determined for all the individual groups/students an aggregate measure of level change can be determined using equation 8. Similarly, an overall slope change can be determined by substituting the individual slope-change coefficients instead on the level change coefficients.

$$\boldsymbol{LC_{overall}} = \frac{\sum_{j=1}^{J} \frac{\mathbf{1}}{\boldsymbol{\sigma_j^2}} \boldsymbol{b_{LC_j}}}{\sum_{j=1}^{J} \frac{\mathbf{1}}{\boldsymbol{\sigma_j^2}}} \quad (\mathbf{8})$$

Where J is the nmber of crews; b_{LC_j} is the level change coefficent estimated for the jth crew; σ_j^2 is the estimated standard error for the jth level change coefficient

Guidelines to conduct a successful MBT study

Preservation of construct validity
Construct validity deals with the legitimacy of using the test or observations to measure a theoretical concept that it intends to evaluate[15]. In other words, construct validity ensures that the measurement scale really measure what it claims to measure. The first step in experimentally testing a hypothesis with MBT is to select the appropriate unit of analysis. For example, if individual student learning achievement is the targeted improvement, the unit of analysis would be each individual student, which would then be aggregated for each time step. Alternatively, if group performance is targeted, learning achievement for the groups would constitute the unit of analysis. It is important to establish a proper unit of analysis and remain consistent with the measurements throughout the longitudinal period because deviations or inconsistencies may yield unsubstantiated conclusions.

The response variable, also referred to as achievement in MBT studies, will vary widely depending on the learning objectives of the exercise. Of course, the learning objectives must be measureable actions that students are able to perform as a result of the educational module. For example, in the author's recent research the learning objective was for participants to be able to identify the safety hazards in planned construction environments. Throughout the longitudinal experiment, the assessment variable remained constant and standard; however, the context of the assignment or the problem to be solved must change. For example during each assessment session, participants were asked to review different construction environments. This could be easily facilitated by using non-repeated randomly selected high resolution photographs of construction environments.

Preservation of Internal Validity
Internal validity is the certainty with which causal inferences can be warranted from the experiment conducted. Threats to internal validity are discussed in several publications. Practices and methods that preserve internal validity are listed in Table 1.

Table 1: Preservation of Internal Validity

Threat to internal validity	Description
History	Use of several baselines and staggered introduction of intervention provides control. Each experimental unit's pre-intervention phase provides additional control [21].
Maturation	Relatively short period of study ensures maturity effects are not significant[53]
Withdrawal reaction	Intervention is not withdrawn in MBT studies [39].
Testing	No feedback of previous performance is provided. Different tests testing the same construct are to be used. Stability in repeated measures in the pre-intervention phase verifies the absence of testing effect [31].
Instrumentation/Reporting	Avoidance of this threat by using stringent evaluation technique and training personals. Repeated measures increase confidence [31].
Regression-to-mean	Longitudinal nature of study limits regression to mean to a single phase [54].
Mortality	Providing contingency through increasing the number of participants/groups [53].

Preservation of External Validity

External validity is concerned with the generalizability of research finding to the target population of subjects and settings (Johnson and Christensen 2010). Every research study needs to include appropriate or representative samples of the target population to be able to make generalizations. MBT studies through inter-subject replication enhance external validity. Also literature suggests whenever necessary to use of clustered groups based on demographics or behavior and to reduce systematic variability [27] through which logical generalizations can be made.

Finally, each study is conducted on a sample of subjects in a specific setting. In order to ensure generalization beyond the sample to the target population, sufficiently large samples [55] must be randomly assigned to the study and the study must be conducted under sufficient different settings [31].

Preservation of Reliability

A measurement value always consists of two parts, the true value and the measurement error. Further, the measurement error consists of systematic error which affects validity, while random error will influence reliability [56].Therefore, measures that are valid have a low degree of systematic error while measurements that are reliable have a low degree of random error. Generally, experiments involving repetitive measures such as the MBT are allow researchers to ensure reliability as measurement error is reduced with repeated measures [57].

Case illustration using MBT Design

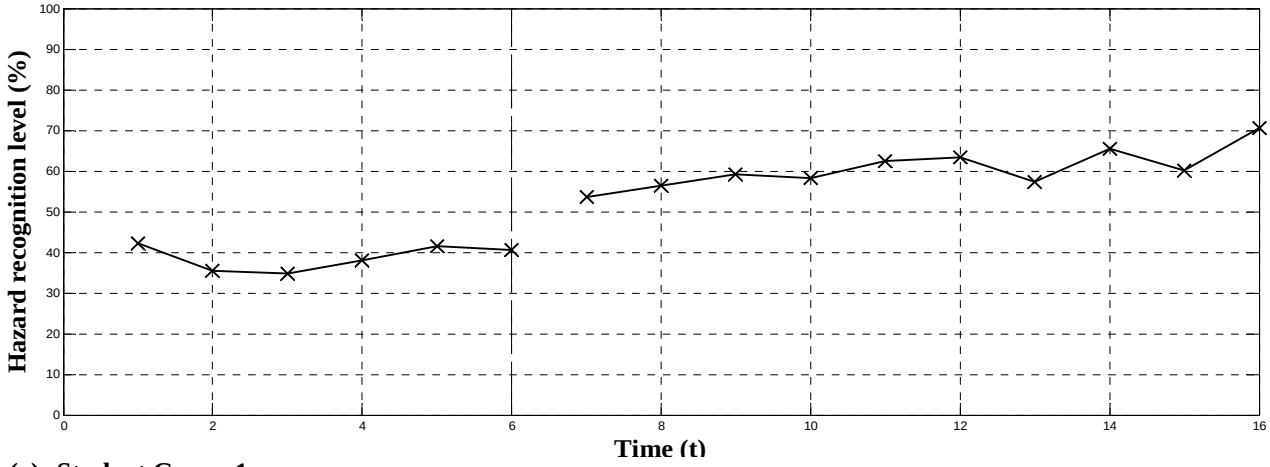
In this section the authors illustrate the application of the MBT design using data that were collected for testing the effects of an intervention. Although more baselines were used to gather reliable and valid data, for illustration and simplicity sake, we present data only for two baselines. The study involved several Student groups that were provided an intervention along with its associated training to improve hazard recognition skills. Table 2 provides a brief description of the experimental elements. Thus, the null hypothesis being tested was: *The*

introduction of the visual cues organized in simple mnemonics will result in higher proportion of hazards recognized in a student population.

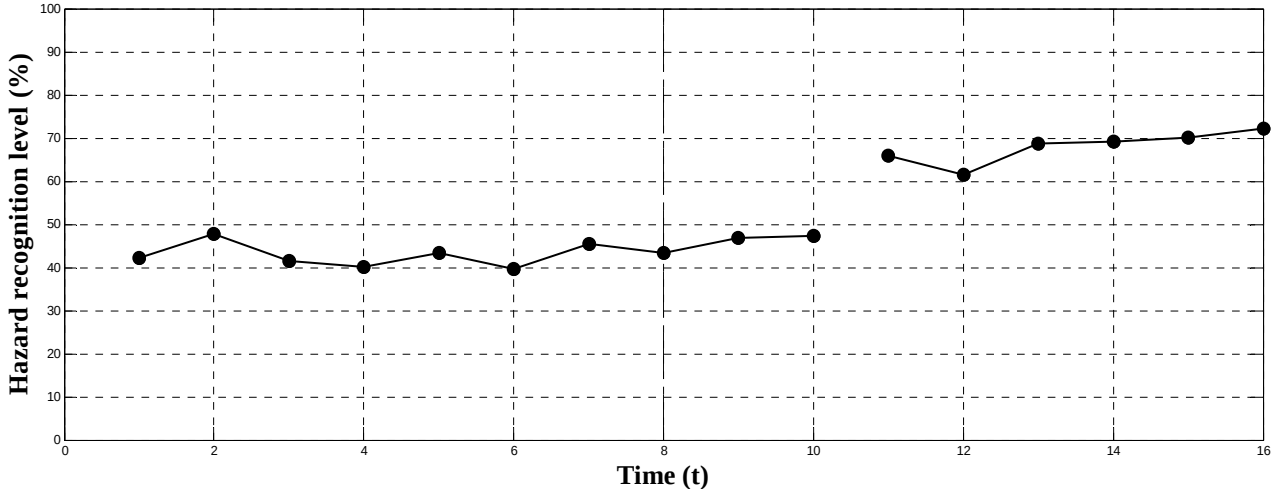
Table 2: Experimental elements for case illustration

Experimental elements	Description
Unit of analysis	Student groups
Learning objective	Identifying hazards in construction environment
Dependent variable	Proportion of hazards identified
Intervention	Visual cues and mnemonics
Testing method	Recognizing hazards different environments and associated tasks

As illustrated previously, baseline measurements were concurrently gathered during the baseline phase and the intervention was introduced in a time-lagged or staggered basis. Figure 2 summarized the proportion of hazards that were identified through time for the two groups in consideration. The analysis results using the method described above are exhibited in Table 4.



(a) Student Group 1



(b) Student Group 2

Figure 2: Results of case illustration

Table 2: Tabulation of results of case illustration

Student Group 1

Reg eqn 37.528+0.352T+14.967D+0.91SC

Predictor	Coef	Std. Error	t	p
β_0	37.53	3.123	12.016	0
β_1	0.352	0.802	0.439	0.669
β_2	14.98	3.694	4.052	0.002
β_3	0.91	0.883	1.03	0.323

Std. Error of est.	3.3549
R-sq	93.50%
R-sq (adj)	91.80%
SS_{Reg}	1928.27
MS_{Res}	11.26

Student Group 2

Reg eqn 41.809+0.357T+18.137D+1.293SC

Predictor	Coef	Std. Error	t	p
β_0	41.81	1.894	22.251	0
β_1	0.357	0.303	1.18	0.261
β_2	18.14	2.737	6.625	0
β_3	1.24	0.724	1.787	0.099

Std. Error of est.	2.75053
R-sq	96.10%
R-sq (adj)	95.20%
SS_{Reg}	2259.38
MS_{Res}	7.57

Step 1: Test of Assumptions for Regression Analysis

The analysis procedure began with the test of the underlying assumptions. The Levene's test for homoscedasticity of error variance considering α=0.05 for the Student Group 1 yielded a p-value of 0.511; while the Anderson-Darling test for normality of errors yielded a p-value of 0.5. Similarly, the underlying assumption of constant variance and normality of the residuals of Student Group 2 were accepted. In the case the underlying assumptions were not achieved, then various transformations provided in most elementary regression textbooks may need to be used to transform the obtained data before further analysis. Considering autocorrelation using the Durban-Watson test indicated no evidence of the presence of autocorrelation.

Step 2: Selection of Mathematical model

Since the ρ=0 was accepted in the previous step, the parameters of Equation 3 (T_t, D_t, and SC_t) and 4 (D_t) were determined by regressing the dependent variable (Y) on the respective predictor variables. Following this, Equation 7 was used to compare the two models based on Equation 3 and 4. The obtained value of F ($F_{obtained}$) was compared with the critical value using a liberal alpha level (α = 0.10) and the respective degrees of freedom (df = 2,12) indicated that the alternate hypothesis was to be accepted. Since the values of β_1 and β_3 were not determined to be 'zero', Equation 3 was appropriate for the observed data. Similarly, Student Group 2 was also appropriately modeled using Equation 3. This implied that both the Student groups exhibited both a level change and a slope change.

From Table 3, we see that Student Group 1 shows a level change (β_2) improvement of 14.98% ($p<0.001$) that occurred just after the intervention was introduced. This level change is the difference between the predicted value of the dependent variable based on the pre-intervention regression line and the post intervention line at the first observational point following the introduction of the intervention. In other words, the projected regression line based on the baseline data for t=7 is 39.99% (b_0+b_1 (T)) and based on the post-intervention data is 54.96%.

The difference between the two phases, then is 14.98% (54.96-39.99%). Similarly, from Table 3, it can be seen that the level change for Student Group 2 was 18.14%

The slope change coefficient from Table 3 for Student Group 1 is 0.91. This value indicates that the value of the slope between the baseline and post-intervention phase changed by 0.91. This means that the slope in the post-intervention phase is equal to the slope in the baseline phase and the observed slope change which is equal to 1.26 (0.352+0.91). This indicates that with each subsequent test, the proportion of hazards recognized increased by 1.26%. Similarly, the Student group 2 exhibited an improvement of 1.65% with every subsequent test.

Step 3: Computation of the Overall Effect size for the MBT study
Finally, the overall level change was computed using Equation 8 based on the reciprocal of error variance, which was computed to be 17.59% (p=0). Similarly, the overall slope change coefficient was determined to be 1.14 (p=0.040). Hence, based on the MBT study the level change (β_2) for the two Student groups, the improvement in the achievement of the stated learning objective can be summarized to be 17.59% with a slope change improvement of 1.14 in the post-intervention phase.

Conclusion

The objective of this paper was to provide a reliable and rigorous methodology for researchers to test the effects of introducing pedagogical interventions. Such methods are essential to understand how various interventions affect student learning and retention. A rationale was provided for adopting the MBT technique for this purpose and a clear protocol to test intervention related research and statistical hypotheses was discussed. At the present time, there is no singular resource for the proper use of the MBT technique in education research. Thus, a rigorous analysis methodology was presented based on which valid inferences can be made. Finally, the discussions were concluded with methods to preserve validity. A case illustration was provided to indicate how the authors have used the MBT method to make valid causal inferences regarding the effects of an intervention to improve student hazard recognition skills.

References

1. Richardson JTE. Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*. 2005;30(4):387-415.
2. Lenehan MC, Dunn R, Ingham J, Signer B. Effects of learning-style intervention on college students' achievement, anxiety, anger, and curiosity. *Journal of College Student Development*. 1994.
3. Wilkins JLM. Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education*. 2004;72(4):331-346.
4. Urdan T. Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *J Educ Psychol*. 2004;96(2):251.
5. Hsieh P, Acee T, Chung WH, et al. Is educational intervention research on the decline? *J Educ Psychol*. 2005;97(4):523.
6. Robinson DH, Levin JR, Thomas GD, Pituch KA, Vaughn S. The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal*. 2007;44(2):400-413.
7. Frazier PA, Tix AP, Barron KE. Testing moderator and mediator effects in counseling psychology research. *Journal of counseling psychology*. 2004;51(1):115.

8. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference.* Belmont: Wadsworth Cengage Learning; 2011.
9. Diggle PJ. *Analysis of longitudinal data.* N.Y.; Oxford: Oxford University Press; 2009.
10. Seethaler PM, Fuchs LS. A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice*. 2005;20(2):98-102.
11. Thompson B, Diamond KE, McWilliam R, Snyder P, Snyder SW. Evaluating the quality of evidence from correlational research for evidence-based practice. *Except Child*. 2005;71(2):181-194.
12. Campbell DT, Stanley JC. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally; 1963.
13. Solomon PL, Cavanaugh MM, Draine J. *Randomized controlled trials : Design and implementation for community-based psychosocial interventions.* New York: Oxford University Press; 2009.
14. Snyder P, Thompson B, Mclean ME, Smith BJ. Examination of quantitative methods used in early intervention research: Linkages with recommended practices. *Journal of Early Intervention*. 2002;25(2):137-150.
15. Kirk RE. *Experimental design : Procedures for the behavioral sciences.* Thousand Oaks: Sage Publications; 2013.
16. Lingard H, Rowlinson SM. *Occupational health and safety in construction project management.* London; New York: Spon Press; 2005.
17. Rossi PH, Lipsey MW, Freeman HE. *Evaluation : A systematic approach.* Thousand Oaks, Calif. [u.a.]: Sage Publ.; 2009.
18. Fennell D, London School of Economics and Political Science. Centre for the Philosophy of the Natural and Social Sciences., London School of Economics and Political Science. Contingency and Dissent in Science Project. *Why and when should we trust our methods of causal inference? : Lessons from james heckman on RCTs and structural models.* London: London School of Economics and Political Science, Contingency and Dissent in Science Project; 2006.
19. Levin JR. Randomized classroom trials on trial. *Empirical methods for evaluating educational interventions*. 2005:3-27.
20. Mosteller F, Boruch RF. *Evidence matters: Randomized trials in education research.* Brookings Inst Press; 2002.
21. Kazdin AE, Kopel SA. On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*. 1975;6(5):601-608.
22. Koehler MJ, Levin JR. Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychol Methods*. 1998;3(2):206.
23. Baer DM, Wolf MM, Risley TR. Some current dimensions of applied behavior analysis. *J Appl Behav Anal*. 1968;1(1):91.
24. LaPointe LL. Sequential treatment of split lists: A case report. *Apraxia of speech: Physiology, acoustics, linguistics, management*. 1984:277-286.
25. Harvey MT, May ME, Kennedy CH. Nonconcurrent multiple baseline designs and the evaluation of educational systems. *Journal of Behavioral Education*. 2004;13(4):267-276.
26. Barlow DH, Hersen M. *Single case experimental designs strategies for studying behavior change.* New York; Oxford; Toronto; Sydney; Paris; Frankfurt [Main i.e.] Kronberg-Taunus: Pergamon Press; 1984.
27. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *Am J Prev Med*. 2007;33(2):162-168.
28. Cooper JO, Heward WL, Heron TE. *Applied behavior analysis.* Upper Saddle River, N.J: Pearson/Merrill-Prentice Hall; 2008.
29. Johnston JM, Pennypacker HS. *Strategies and tactics of behavioral research.* New York: Routledge; 2009.
30. Richards S. *Single subject research : Applications in educational and clinical settings.* San Diego: Singular Pub. Group; 1999.
31. Janosky JE. *Single subject designs in biomedicine.* Dordrecht: Springer; 2009.
32. Borman GD. Experiments for educational evaluation and improvement. *Peabody Journal of Education*. 2002;77(4):7-27.
33. White KR, Pezzino J. Ethical, practical, and scientific considerations of randomized experiments in early childhood special education. *Topics in Early Childhood Special Education*. 1986;6(3):100-116.
34. Campbell DT, Stanley JC, Gage NL. *Experimental and quasi-experimental designs for research.* Chicago, Ill.: R. McNally; 1966.
35. Gast DL. *Single subject research methodology in behavioral sciences.* Routledge New York, NY; 2010.

36. Singer JD, Willett JB. *Applied longitudinal data analysis : Modeling change and event occurrence.* Oxford; New York: Oxford University Press; 2003.
37. Willett JB. Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*. 1989;49(3):587-602.
38. Beins B. *Research methods : A tool for life.* Boston: Pearson/A&B; 2009.
39. Barlow DH, Nock M, Hersen M. *Single case experimental designs : Strategies for studying behavior for change.* Boston: Pearson/Allyn and Bacon; 2009.
40. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prevention Science*. 2000;1(1):31-49.
41. Freeman KA. Treating bedtime resistance with the bedtime pass: A systematic replication and component analysis with 3-year-olds. *J Appl Behav Anal*. 2006;39(4):423.
42. Tincani M, Crozier S, Alazetta L. The picture exchange communication system: Effects on manding and speech development for school-aged children with autism. *Education and Training in Developmental Disabilities*. 2006;41(2):177.
43. Watson P, Workman EA. The non-concurrent multiple baseline across-individuals design: An extension of the traditional multiple baseline design. *J Behav Ther Exp Psychiatry*. 1981;12(3):257-259.
44. Gliner JA, Morgan GA, Leech NL. *Research methods in applied settings : An integrated approach to design and analysis.* London: Routledge; 2009.
45. Davis CS. Statistical methods for the analysis of repeated measurements. . Updated 2002.
46. Hayes SC. Single-case research designs: Methods for clinical and applied settings. *Journal of Behavior Therapy and Experimental Psychiatry Journal of Behavior Therapy and Experimental Psychiatry*. 1983;14(1):81.
47. Huitema BE. *The analysis of covariance and alternatives : Statistical methods for experiments, quasi-experiments, and single-case studies.* Hoboken, N.J.: Wiley; 2011.
48. Hoogeboom TJ, Kwakkenbos L, Rietveld L, den Broeder AA, de Bie RA, van den Ende CHM. Feasibility and potential effectiveness of a non-pharmacological multidisciplinary care programme for persons with generalised osteoarthritis: A randomised, multiple-baseline single-case study. *BMJ open*. 2012;2(4).
49. Nourbakhsh MR, Ottenbacher KJ. The statistical analysis of single-subject data: A comparative examination. *Phys Ther*. 1994;74(8):768-776.
50. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *Int J Technol Assess Health Care*. 2003;19(04):613-623.
51. Huitema BE, Mckean JW. Design specification issues in time-series intervention models. *Educational and Psychological Measurement*. 2000;60(1):38-58.
52. Huitema BE, McKean JW. Identifying autocorrelation generated by various error processes in interrupted time-series regression designs A comparison of AR1 and portmanteau tests. *Educational and psychological measurement*. 2007;67(3):447-459.
53. Lingard H. The effect of first aid training on objective safety behaviour in australian small business construction firms. *Construction Management & Economics*. 2001;19(6):611-618.
54. Johnson B, Christensen LB. *Educational research : Quantitative, qualitative, and mixed approaches.* Thousand Oaks, Calif.: SAGE Publications; 2011.
55. Morgan DL, Morgan RK. *Single-case research methods for the behavioral and health sciences.* Los Angeles: SAGE; 2009.
56. Robson LS, National Institute for Occupational Safety and Health., Institute for Work & Health. Guide to evaluating the effectiveness of strategies for preventing work injuries how to show whether a safety intervention really works. . Updated 2001.
57. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. . Updated 2012.