

BEYOND THE FOUR LEVELS: AN EVALUATION MODEL FOR GROWTH AND SUSTAINABILITY

by

DARRYL GARDNER

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: EVALUATION AND RESEARCH

Approved By:

Advisor

Date

**© COPYRIGHT BY
DARRYL GARDNER
2019
All Rights Reserved**

DEDICATION

This work is dedicated in loving memory of Jimmie and Hazel Knight.

ACKNOWLEDGMENTS

Thanks Mom, Cameron, Mylitta, LaShantinette, Carmen, Derek, Myra, Marcus, Marvin, Carlos,
Thomas, Dorthy, Jeffery, and Lammar.

Thanks to my dissertation committee: Prof. Shlomo Sawilowsky, Prof. D. Carl Freeman, Prof.
Ingrid Guerra-Lopez and Prof. Barry Markman.

Thanks to the Finish Line.

Special thanks to Orchards Children's Services.

TABLE OF CONTENTS

Dedication.....	ii
Acknowledgements.....	iii
List of Tables.....	vi
List of Figures.....	vii
Chapter 1: Introduction.....	1
Background.....	1
Problem Statement.....	2
Purpose of Study.....	3
Conceptual Framework.....	4
Research Questions.....	4
Limitations.....	4
Key Terms and Definitions.....	5
Chapter 2: Literature Review.....	7
Objectives-based Evaluation.....	7
Post Objectives-based Evaluation.....	10
Kirkpatrick’s Four Level Model.....	13
Building on Kirkpatrick.....	17
Chapter 3: Methodology.....	20
Development of Proposed Evaluation Model.....	20
Research Design.....	21
Meta Evaluation.....	24
Chapter 4: Results.....	27

Operationalization of Proposed Model.....	27
Stage 1: Reaction.....	28
Stage 2: Retention.....	29
Stage 3: Transfer.....	31
Stage 4: Impact.....	33
Meta Evaluation.....	34
Chapter 5: Discussion	
Proposed Model Fit in Evaluation Research and Theory.....	38
Key Findings.....	40
Meta Evaluation Findings.....	42
Limitations.....	45
Recommendations for Future Research.....	46
Implications for Practice.....	46
Implications Beyond Evaluation and Research Theory.....	47
Summary and Conclusion.....	49
Abstract.....	55
Autobiographical Statement.....	56

LIST OF TABLES

Table 1: Frequency of Kirkpatrick Level Usage by eLearning Guild.....	16
Table 2: Frequency of Kirkpatrick Level Usage by ASTD.....	17
Table 3: Satisfaction Survey Results.....	28
Table 4: Group Statistics: Program Participants and Students Regularly Enrolled.....	32
Table 5: T-Test Table: Program Participants and Students Regularly Enrolled.....	32
Table 6: Utility Ratings.....	34
Table 7: Feasibility Ratings.....	34
Table 8: Propriety Ratings.....	35
Table 9: Accuracy Ratings.....	35
Table 10: Evaluation Accountability Ratings.....	36
Table 11: Relevant Metaevaluation Standards for Evaluation Designs.....	43

LIST OF FIGURES

Figure 1: Kirkpatrick Evaluation Model.....	14
Figure 2: Proposed Evaluation Model.....	21
Figure 3: Exam Averages by Semester.....	30
Figure 4: Homework Completion Averages.....	30
Figure 5: Average Class Attendance By Semester.....	31
Figure 6: Grade Distributions: Group A vs Group B (Fall 2012-Winter 2015).....	33
Figure 7: Proposed Model: Reimagined.....	37

CHAPTER 1: INTRODUCTION

Background

Fitzpatrick, Sanders and Worthen (2011) defined evaluation as the identification, clarification and application of defensible criteria to determine an evaluation object's value in relation to those criteria. They asserted that all programs, to be effective, must have the element of evaluation. Until late 1940's, evaluation was synonymous with testing (Hogan, 2007). The purpose of evaluation, in its early conception, was to measure individual differences (Hogan 2007). After the 1957 Russian launch of Sputnik prompted a key shift in curriculum development in the United States, it became apparent that evaluation instruments needed to provide information about the relative effectiveness of courses, training and programs. Objectives-based evaluation was no longer exclusively valid as it related to measuring efficiency and outcomes. Innovation and global competition inspired innovation in evaluation. A need for continual assessment and data collection became extremely important for curriculum developers, teachers and decision makers (Taylor & Crowley, 1972).

Kirkpatrick (1959) introduced a framework for evaluating training programs in a series of four articles to the American Society for Training and Development (ASTD). The Kirkpatrick model has become a standard in organizing evaluations across many industries, such as in business, government, education and the military (Watkins et al, 1998; Reio et al, 2017). Often referred to as the four-level model, Kirkpatrick outlined the evaluation process in four steps: (1) reaction, (2) learning, (3) behavior and (4) results.

At Level 1, participants' reactions to the program are assessed. Kirkpatrick (1959) discussed reaction from the perspective of how well the participants liked the program. During Level 2, learning, quantifiable indicators of learning are measured. At Level 3, behavior outcomes

are examined to determine the extent to which knowledge is transferred and applied on the job. At Level 4, outcomes are intended to determine the broader impact the program or training in regard to the organizational goals and objectives. According to Gunderman & Chan (2015):

Proponents of Kirkpatrick's approach hold that, properly applied, his model can enrich education both prospectively in the planning phase and retrospectively in the evaluation phase. Although it has been for more than 50 years after its initial publication, this approach has remained the dominant “model” in educational evaluation, in spite of the fact that Kirkpatrick's original publication construed the four criteria for learning evaluation as individualized techniques for conducting the evaluation, rather than as a hierarchal model. For more than 50 years, this model has evolved as the educational training industry has expanded, especially in the corporate world. (p. 1323)

Performance measurements and outcome assessments are expanding world-wide (Fitzpatrick, Sanders & Worthen, 2004). Moreover, evaluation has become a central force within organizations across many different industries. Program evaluation is changing in a variety of ways to help stakeholders obtain vital information in order to make important decisions. These decisions can range from scaling up the efforts of a particular program to terminating the program all together.

Due to the changing grant requirements within the Department of Education and various other funding and accreditation entities, program evaluation in education has grown in popularity over the years (Booker, 2016). The need to holistically (and efficiently) evaluate educational programs is critical in the 21st century.

Problem Statement

Kirkpatrick (1959) offered a multilevel approach to accomplish such a task, however, the framework falls short in critical areas. The model was criticized for its inability to connect programmatic outcomes to programmatic activities. The model does not provide information on why a program is effective (Holton, 1996; Bates 2004). Additionally, as Pulichino (2007) and Kennedy (2012) pointed out, Kirkpatrick didn't provide guidelines on how to move through the levels and many who use the model fail to evaluate at Level 3 and Level 4. Level 3 and Level 4 potentially provide the most information in regard to overall institutional impact (Pulichino, 2007).

Purpose of the Study

Given the limitations in the Kirkpatrick (1959) evaluation model, the purpose of this study is to develop an evaluation model that can be used to evaluate educational programs using prospective or retrospective data. Kirkpatrick's original framework will be used to develop a new evaluation model that will take into account contextual and process factors. The new model will be able to address the major flaw of the Kirkpatrick's model by providing causal linkages between programmatic activities/functions and programmatic outcomes. The new model will also be able to identify key programmatic functions that are essential to replicating and/or scaling up successful programs.

A developmental mathematics program at an urban, research university in the Midwest will serve as the backdrop for this study. Archived data that has been collected at the institutional, departmental and programmatic levels will be used. The evaluation of the developmental mathematics program is to determine the effectiveness of the model in regard to directly connecting programmatic activities with outcomes by examining contextual and process factors. After the preliminary examination of the model, a meta-evaluation of the evaluation model will be conducted in accordance to the Joint Committee's Program Evaluation Standards (1994).

Conceptual Framework

The new evaluation model will be based on the program evaluation standards developed by the Joint Committee on Standards for Educational Evaluation (1994). The four attributes of sound educational evaluation practices defined by the Joint Committee are (1) utility, (2) feasibility, (3) propriety and (4) accuracy. It will incorporate Dubin's (1996) six characteristics or components of models: (1) elements or units – represented as constructs – are the subject matter, (2) there are relationships between the constructs, (3) there are boundaries or limits of generalization, (4) system states and changes are described, (5) deductions about the theory in operation are expressed as propositions or hypothesis, and (6) predictions are made about units.

Hill's (1986) theory on the functions of evaluation will be used as a guiding factor in the development of the new model. Hill (1986) outlined four functions of evaluation: (1) evaluation for accountability, (2) evaluation for intervention, (3) evaluation for decision making and (4) evaluation for meaning, where these functions are not mutually exclusive, and they should overlap.

Research Questions

The research questions that will be examined in this study are as follows:

Research Question 1: How effective is the new model for detecting causal linkages between programmatic activities/functions and outcomes?

Research Question 2: How well does the new model meet the Joint Committee's Program Evaluation Standard requirements for utility, feasibility, propriety, and accuracy?

Research Question 3: Where does the new model fit in the body of evaluation research and theory?

Limitations

There are potential limitations associated with this study. The efficacy of the evaluation instrument developed in this study is being evaluated using one developmental mathematics

program at a large, urban research university. Thus, any conclusions drawn may not be able to be generalized to similar programs or institutions.

Key Terms and Definitions

Evaluation – The formal determination of quality, effectiveness, or value of a program (Worthen & Sanders, 1987).

Evaluation Standard – “A principle mutually agreed to by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation.” (Joint Committee on Standards for Educational Evaluation, 1994, p. 3)

Meta-evaluation - A systematic review of an evaluation in order to define the quality of the methods and results of the evaluation (Cooksy & Caracelli, 2009).

Model – A practical and systematic guide in planning and execution of implementing a theoretical process – usually done in steps of phases (Nilsen, 2015). Used interchangeably in this study with framework.

Taxonomy – A scheme of classification (Holton, 1996)

Stakeholder - Individuals or groups that may be involved in or affected by a program evaluation (Joint Committee on Standards for Educational Evaluation, 1994).

Reaction: refers to how well the participants liked the program (Kirkpatrick, 1959a).

Learning: refers to the principles, facts and techniques that were learned (Kirkpatrick, 1959b). Used interchangeably in this study with retention.

Behavior: refers to the changes in behavior that resulted from the program – how the learning was applied (Kirkpatrick, 1960a). Used interchangeably in this study with transfer.

Results: refers to the tangible results of the program in terms of reduced cost, improved quality, improved retention, etc. (Kirkpatrick, 1960b). Used interchangeably in this study with impact.

Conceptual Factors: Refer to the environment or ecology of the program's implementation.

Process Factors: Refer to "how" in regard to the program's implementation.

CHAPTER 2: LITERATURE REVIEW

Objectives-based Evaluation

Objectives-based evaluation, also referred to as objectives-oriented, is a classification of evaluation approaches that focuses on the specification of goals/objectives and the measurement of outcomes. According to Stufflebeam (2001), the objectives-based study has been the most prevalent approach in program evaluation. It has common-sense appeal and program administrators have had a great amount of experience with it.

The objectives-based approach is especially applicable in assessing tightly focused projects that have clear, supportable objectives. Even then, such studies can be strengthened by judging project objectives against the intended beneficiaries' assessed needs, searching for side effects, and studying the process as well as the outcomes. (p. 17)

Common criticisms of objectives-based approach are that such studies do not provide timely nor pertinent information to improve a program's process. According to Stufflebeam (2003), evaluation should be done with the intention of not to prove, but rather improve. Moreover, often times the information often is far too narrow to constitute a sufficient basis for judging the program's merit and may credit unworthy objectives (Scriven, 1972; Stufflebeam, 2001; Fitzpatrick, Sanders & Worthen, 2004)

The Tylerian Evaluation Approach

Tyler (1942) was credited for the conceptualization of objectives-based evaluation, based on the Eight Year Study (Smith & Tyler, 1942). It was a comprehensive, longitudinal study that examined 30 schools through four years of secondary school and four years of college. According to Madaus and Stufflebeam (1988), Tyler's work on the Eight Year Study was the best available description of how evaluators and teachers can work cooperatively to clarify instructional

objectives and develop indicators of students' continuous progress toward the mastery of a whole range of learning outcomes.

Tyler's (1942) process for determining the extent to which program objectives are being met can be outlined in the following steps: (1) Establish broad goals and objectives, (2) Classify the goals or objectives, (3) Define objectives in behavioral terms, (4) Find situations in which achievement of objectives can be shown, (5) Develop or select measurement techniques, (6) Collect performance data, and (7) Compare performance data with behaviorally stated objectives (Fitzpatrick, Sanders & Worthen, 2004). Tyler's framework laid the foundation for evaluators interested in linking program objectives to measured outcomes. This would shift educational evaluation for nearly half a century (Christie & Alkin, 2013).

Tyler's framework has several advantages. It is systematic and easy to apply. Moreover, it allows curriculum developers to focus on strengths and weaknesses relative to the program, rather than being solely concerned about individual performance. However, there are also several weaknesses associated with Tyler's model. The model fails to suggest, specifically, how individual objectives should be met. Moreover, it does not provide standards or suggest how standards should be met or addressed (Guba & Lincoln, 1981). As some evaluation theorists pointed out (Sriven, 1972; Fitzpatrick, Sanders & Worthen, 2004), Tyler's emphasis on predetermined objectives restricts creativity in program development and there is also a disregard for formative assessment.

Provus' Discrepancy Model

In the tradition of Tyler, Provus (1971) developed the Discrepancy Evaluation Model. Provus viewed evaluation as the comparison of a desired standard to actual performance. Like Tyler, Provus outlined a process for evaluation based on standards or objectives (Fox, 2011). This process can be summarized by the following steps: (1) Agreeing on standards (objectives), (2)

Determining whether a discrepancy exists between the performance of some aspect of a program and the standards set for the performance, and (3) Using information about discrepancies to decide whether to improve, maintain, or terminate the program or some aspect of it (Fitzpatrick, Sanders & Worthen, 2004). Many view his model as a problem-solving set of procedures that seeks to identify weaknesses. According to McKenna (1981):

Provus considers discrepancies to be the essential clue in program evaluation. Discrepancies point out differences that exist between what program planners think is happening in the program and what's actually happening. Provus recommends that when discrepancies occur, either program performance or program design standards be changed.
(p. 10)

Fitzpatrick, Sanders and Worthen (2004) summarized Provus' theory on the program development and evaluation in five stages:

1. Definition – During the definition stage, the focus is on defining programmatic goals, processes, and/or activities.
2. Installation – During this stage, the program's definition is used as the standard to judge program operation. A series of congruency tests are performed to identify any discrepancies between expected and actual implementation of the program or activity.
3. Process – During the process stage, data is gathered on the progress of participants to determine whether behaviors changed as expected.
4. Product – At the product stage, the evaluation determines whether the terminal objectives of the program have been met.
5. Cost-benefit analysis – During this optional stage, a cost analysis of the program is compared to the cost of analyses of comparable programs.

Provus' Discrepancy approach was designed to evaluate and facilitate program development in a large public school system (Provus, 1971). Due to its complex nature, the model works best in large, complex organizations. However, it is still vulnerable to the same criticisms of similar objective-based models. Scriven (1972) cautioned against objectives-based evaluation without also evaluating the quality of the objectives/goals. If goals are not worth achieving, then it is uninteresting, and essentially useless, to know how well they are achieved.

Post Objectives-Based Models

The need to expand the evaluation process beyond the objectives-based framework became essential also due to global competition. There have been several models that have sought to address flaws in the Tylerian approach. Four such models that have dominated the literature are Stufflebeam's Context, Input, Process and Product (CIPP) model, Scriven's Goal-Free model, Stake's Responsive model and Eisner's Connoisseurship model.

CIPP Model

Stufflebeam introduced the CIPP model in 1966. Designed to service the needs of administrators and decision makers, it is still widely used today (Mertens & Wilson, 2012; Zhang et al., 2011; Warju, 2016). The key components/concepts of the model are context evaluation, input evaluation, process evaluation and product evaluation.

At the context evaluation phase of the CIPP model, the evaluator identifies the needs, assets, and resources (Mertens & Wilson, 2012). To develop this context, the evaluator compiles and assesses background information, and solicits feedback from key stakeholders and leadership. During the input evaluation phase, information is collected regarding the mission, goals, and plan of the program in order to assess the program's strategy and merit. According to Fitzpatrick et al. (2011), the intent of this stage is to choose an appropriate strategy to implement to resolve any

programmatic issues. The third phase of the CIPP model is process evaluation. During process evaluation, the quality of the implementation of the program is investigated. The primary objectives of this stage are to provide feedback regarding the extent to which planned activities are carried out, guide staff on how to modify and improve the program plan and assess the degree to which participants can carry out their roles (Stufflebeam, 2003). The final component to CIPP is product evaluation. During this stage, the positive and negative effects the program had on its target audience and key stakeholders are examined (Mertens & Wilson, 2012). Moreover, judgments of stakeholders and relevant experts are analyzed. Outcomes that impact the group, subgroups and individuals are also viewed (Stufflebeam, 2003).

The CIPP model has several obvious advantages. With an emphasis on decision making, this model provides administrators with a more complete process in regard to evaluating a program. Unlike the Tyler model, CIPP has built-in features that include formative assessment. However, there are also drawbacks associated with Stufflebeam's model. Primarily, CIPP is time consuming and can be very costly.

Responsive Model

Stake (1975) introduced this Responsive model, evaluation is based on what people do naturally to evaluate things - observe and react (Stake, 1975). The model has three components which are antecedent, transaction and outcome (Thanabalan et al., 2015). In responsive evaluation, stakeholders actively participate in the entire process. They are involved in the formulation of questions, the selection of participants and the interpretation of results. Moreover, Stake emphasized that evaluation should be conducted where the learning occurs. The goal is improved communication, so the model responded to emerging issues learned through interaction and observation instead of giving too much attention to predetermined issues (Stake, 1975).

Stake's model is strong in the fact that it is sensitive to all involved. Responsive evaluators work to understand values, beliefs and cultures of all stakeholders. Moreover, it is extremely flexible and has the ability to incorporate a variety of methodologies. Many theorists concur that the biggest weakness of Stake's approach, however, is the fact that the evaluator is subject to client manipulation (Thanabalan et al., 2015).

Goal-Free Model

Michael Scriven (1972) developed a goal-free evaluation model. It was introduced as a formative and summative evaluation process that advocates gathering data on a broad array of actual effects and evaluating the importance of these effects in meeting demonstrated needs (Patton, 1997). The purpose of goal-free evaluation was to gain knowledge and develop, maintain, or modify an implemented program to meet the needs of participants through evaluation of side effects.

Goal-free evaluation was needs-based which stands in contrast to the traditional goal-based evaluation models discussed earlier. Scriven (1972) believed that goals create a tunnel vision and that prohibits the evaluator from seeing the positive and negative side effects of a program. Another criticism he addressed the idea that stated goals, which are often unrealistic, are rarely the same as the attained goals. Thus, by removing the goal focus, the external evaluator can focus on the actual effects of program and the evaluation is not affected by the shifting goals (Cole, 2015).

There are advantages in taking a goal-free approach to evaluation; however, it is important to note that Scriven does not believe goal-free evaluation works alone. Instead, he pairs it with a quantitative, goal-based evaluation completed by an internal evaluator while an external evaluator focuses on the qualitative, goal-free portion (Cole, 2015). This reduces overlap in assessment of a

program and frees the external evaluator to assess, without bias, if the needs of students are being met (Patton, 1997).

Connoisseurship Model

Eisner (1979) developed a connoisseurship approach to evaluation. It was built on two closely related concepts: connoisseurship and criticism. Connoisseurship was defined as the art perception that makes appreciation of complexity possible (Dunin-Woyseth & Nilsson, 2013). Criticism, according to Eisner, is the art of disclosing qualities of an entity that connoisseurship perceives (Eisner, 1979). According to Smith (2005):

In such a disclosure, the educational critic is more likely to use what Eisner calls “nondiscursive”—a language that is metaphorical, connotative, and symbolic. It uses linguistic forms to present, rather than represent, conception or feeling. Educational criticism, in Eisner’s formulation, has three aspects. The descriptive aspect is an attempt to characterize and portray the relevant qualities of educational life—the rules, the regularities, the underlying architecture. The interpretive aspect uses ideas from the social sciences to explore meanings and develop alternative explanations—to explicate social phenomena. The evaluative aspect makes judgments to improve the educational processes and provides grounds for the value choices made so that others might better disagree (n.p.).

Eisner's model stands in contrast to the traditional scientific models. Its critics have faulted it for its lack of methodological rigor and have also argued that use of the model requires a great deal of expertise (Smith, 2005).

Kirkpatrick’s Four Level Model

Kirkpatrick’s model is a multi-level evaluation approach with usefulness across a wide range of industries (Watkins et al, 1998, see Figure 1). At Level 1, reaction, how well participants

liked the program is evaluated (Kirkpatrick, 1959a). Reaction is also commonly referred to as a "smile sheet." Kirkpatrick (1959a) emphasized not to include measurement of any learning at this phase. "In practice, measures at this level have evolved and are most commonly directed at assessing trainees' affective responses to the quality (e.g. satisfaction with the instructor) or the relevance (e.g. work-related utility) of training" (Bates, 2004, p. 341). Participants are able to give direct feedback and make comments/suggestions on how training can be improved in the future. According to Kirkpatrick (1994), all programs should be evaluated at this level due to the fact that it is quick, easy and generally inexpensive. Additionally, Kirkpatrick believed the reaction of the participants have important consequences for Level 2 of this model (learning). He asserted that a positive reaction does not guarantee learning, but a negative reaction almost certainly reduces its possibility (Kirkpatrick, 1994)

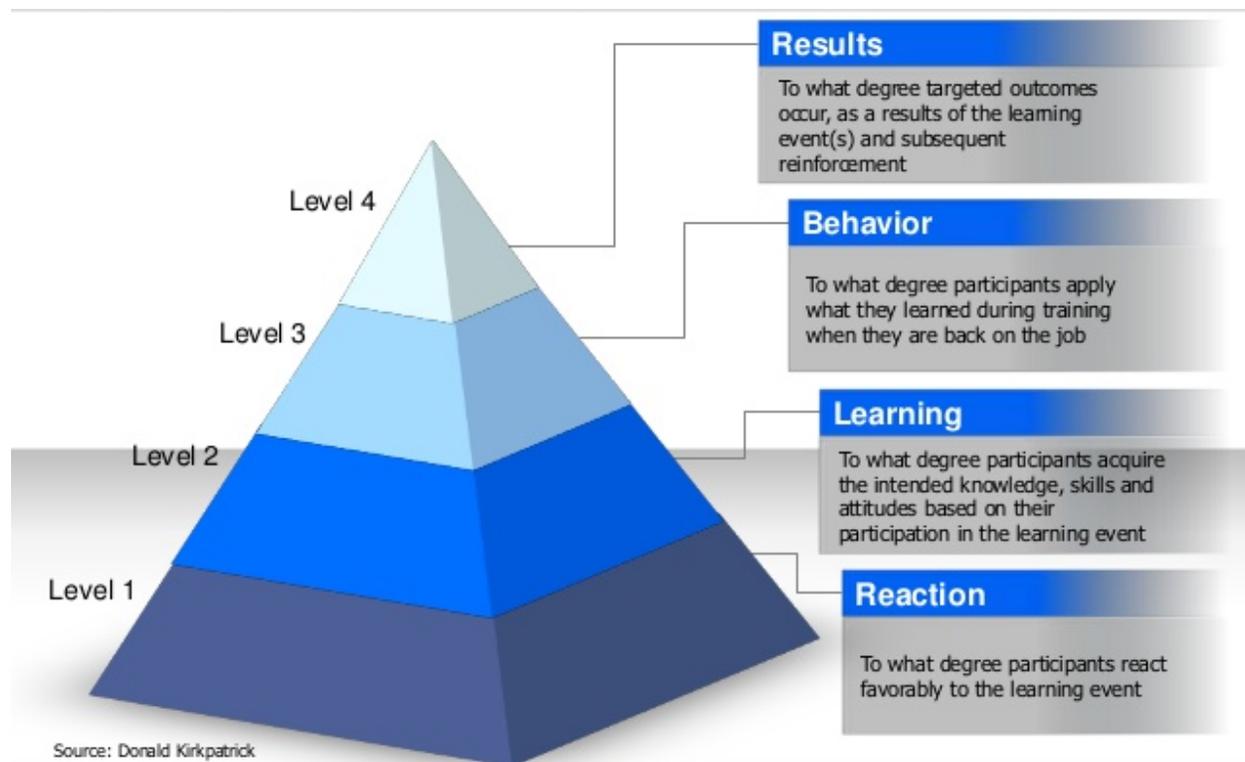


Figure 1. Kirkpatrick Evaluation Model (source: xtlearn.net)

At Level 2, learning was measured. Learning was defined by what principles, facts and techniques were absorbed by participants (Kirkpatrick, 1959b). Moreover, learning should be measured shortly after training/coursework ends. Kirkpatrick emphasized that evaluators should not be concerned with the application of these principles, facts and techniques at this level. Application is measured at the third level (behavior).

Behavior is the third level of Kirkpatrick's model. At Level 3, the transfer that has occurred in learners' behavior due the training or instruction is measured (Kirkpatrick, 1994). Behavior was typically measured by conducting observations and interviews over time. For trainers/teachers, outcomes at this level provided the most useful information in terms of instructional effectiveness.

Level 4, results, refers to the tangible results of the program in terms of reduced cost, improved quality, improved retention, etc.

Despite the widespread use of Kirkpatrick's framework, it was criticized because the model fails to provide evidence for causal linkages between training (or programmatic) outcomes. It only focuses on the how and fails to inform the evaluation audience of why outcomes are reached. Thus, Alliger and Janak (1989) asserted that Kirkpatrick's framework was best labeled as a taxonomy. "Attempts to test causal assumptions within a taxonomy are futile because, by definition, taxonomies classify rather than define causal constructs" (Holton, 1996, pg. 6). Holton (1996) suggested that, causal conclusions are a necessary part of evaluation and a more complete model is required. "No evaluation model can be validated without measuring and accounting for the effects of intervening variables" (Holton, 1996, pg. 7).

Dubin (1976) and later Klimoski (1991), identified six components that models should possess. Holton (1996) summarized the components:

1. Elements or units – represented as constructs – are the subject matter.

2. There are relationships between the constructs.
3. There are boundaries or limits of generalization.
4. System states and changes are described.
5. Deductions about the theory in operation are expressed as propositions or hypothesis.
6. Predictions are made about units

Holton (1996) pointed out Kirkpatrick’s four levels do not meet any of the above criteria to be designated a model. Holton (1996) proposed “an integrative evaluation model that accounts for the impact of the primary and secondary intervening variables” (Holton, 1996, p. 7). However, no progress was given after this initial step in establishing such a model.

Another important criticism is how infrequent the model is completely implemented. More specifically, the model is irregularly used at Level 3 and Level 4. Pulichino (2007) conducted an extensive study to examine the usage of the levels. A survey was administered to 446 members of the eLearning Guild, an organization of training professionals. The results were summarized by Kennedy (2012) and are provided in Table 1.

Table 1.
Frequency of Kirkpatrick Level Usage by eLearning Guild (Pulichino, 2007)

Kirkpatrick Level	Frequency of Usage				
	Never	Rarely	Sometimes	Frequently	Always
1. Reaction	3.1	5.2	7.2	23.3	61.2
2. Learning	4.0	11.0	28.0	37.0	19.1
3. Behavior	13.0	34.6	32.5	14.1	5.8
4. Results	33.0	40.5	12.8	9.4	4.3

About 92% of the respondents conducted Level 1 evaluations and about 84% conducted Level 2 evaluations. There was a significant drop off in regard to evaluations conducted at Level 3 and Level 4. At Level 3, about 48% of the respondents rarely or never conducted evaluations. About 74% rarely or never conducted evaluations at Level 4. Although fewer evaluations were

conducted at Level 3 and Level 4, Pulichino (2007) also reported the majority found value in both levels. Regarding evaluating the effectiveness of training programs, about 97% found Level 3 evaluations to be valuable and 97% found Level 4 to be valuable. Similarly, about 95% found Level 3 to be valuable in evaluating desired change in job performance and 97% found Level 4 to be valuable in regard to evaluating desired organizational results.

The American Society for Training and Development conducted a similar study in 2009. ASTD surveyed 704 of its members to determine the extent in to which they implement Kirkpatrick's level. The results are presented below in Table 2.

Table 2.
Frequency of Kirkpatrick Level Usage
by ASTD (Kennedy, 2012)

Kirkpatrick Level	Frequency of Usage
	Percentage
1. Reaction	91.6
2. Learning	80.8
3. Behavior	54.6
4. Results	36.9

The ASTD also found there was high usage of Level 1 and Level 2, about 92% and 81%, respectively. However, there was also a significant drop off at Level 3 and Level 4, about 55% and 37% respectively.

Pulichino (2007) also examined the barriers to conducting evaluations at Level 3 and Level 4. Although evaluators placed high value on those levels, they also identified what prevented evaluations from being conducted. For Level 3, over 83% of the respondents identified difficulty in accessing the data required, lack of management support and time-consumption as barriers. Those were also the key barriers at Level 4 along with cost.

Building on Kirkpatrick

Theoretical Framework

The goal of this study is to create a new evaluation model based on Kirkpatrick's original framework in order to evaluate educational programs and detect causal linkages between programmatic activities and outcomes. The program evaluation standards developed by the Joint Committee on Standards for Educational Evaluation (1994) will be used as the theoretical framework for the new model. The Joint Committee outlined a set of 30 standards that are designed to help evaluators and consumers alike in judging the quality of a particular evaluation (Fitzpatrick, Sanders & Worthen, 2011). The 30 standards are grouped in five categories: (1) Utility Standards which help to assure that stakeholders find program evaluation processes valuable to their needs, (2) Feasibility Standards which are intended to increase evaluation effectiveness and efficiency, (3) Proprietary Standards which support legal, fair and just evaluations (4) Accuracy Standards which are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality, and (5) Evaluation Accountability Standards which encourage adequate documentation of evaluations and a meta-evaluative perspective focused on improvement and accountability for evaluation processes and products (Joint Committee on Standards for Educational Evaluation, 1994; Booker, 2016).

Building on Kirkpatrick's framework, the model will incorporate Kirkpatrick's four stages – reaction, retention (learning), transfer (behavior) and impact (results). However, the major distinction is that the new model will not be multilevel in nature. All stages will be interrelated and may be revisited at any time of the evaluation. In other words, the new model will give space for formative assessment. Additionally, unlike Kirkpatrick, evaluating organizational/institutional impact will be fundamental to the new model. Impact will serve as the backdrop of the entire evaluation process and will be considered during and in-between stages. Lastly, contextual and

process factors will be examined on a continuous basis throughout each stage in order to connect key programmatic components/activities to measured outcomes.

CHAPTER 3: METHODOLOGY

Introduction

The purpose of this study is to assess the efficacy of a proposed evaluation model as a tool for evaluating educational programs and detecting causal linkages between programmatic activities and outcomes. A meta-evaluation of the model, based on the 2011 Joint Committee's Program Evaluation Standards, will also be conducted. The research questions that will be examined in this study are as follows:

Research Question 1: How effective is the new model for detecting causal linkages between programmatic activities/functions and outcomes?

Research Question 2: How well does the new model meet the Joint Committee's Program Evaluation Standard requirements for utility, feasibility, propriety, and accuracy?

Research Question 3: Where does the new model fit in the body of evaluation research and theory?

The methodology will be divided into three sections: (1) Development of the proposed evaluation model, (2) Research Design and (3) Meta-evaluation.

Development of Proposed Evaluation Model

Development Procedures

Research related to the limitations of the Kirkpatrick Model was conducted (and outlined in the literature review) in effort to develop the proposed evaluation model. The limitations of the Kirkpatrick model that the new model will address are as follows: lack of causal linkages between programmatic activities and outcomes and (2) the lack of model features (taxonomy).

Building on Kirkpatrick's framework, the proposed evaluation model will incorporate four stages – reaction, retention (learning), transfer (behavior) and impact (results). However, the major

distinction is that the proposed model is not hierarchal in nature. All stages are interrelated and may be revisited at any time of the evaluation. Moreover, unlike Kirkpatrick, evaluating organizational/institutional impact is fundamental to the proposed model. Impact serves as the backdrop of the entire evaluation process and is considered during and in-between stages. Lastly, contextual and process factors are examined on a continuous basis throughout each stage in order to connect key programmatic components/activities to measured outcomes. There are four contextual factors defined within the proposed model: (1) history, (2) structure, (3) policies and (4) resources. Additionally, there are four process factors defined within the proposed evaluation model: (1) collaboration, (2) communication, (3) community and (4) data collection. The proposed evaluation model can be found in Figure 2.

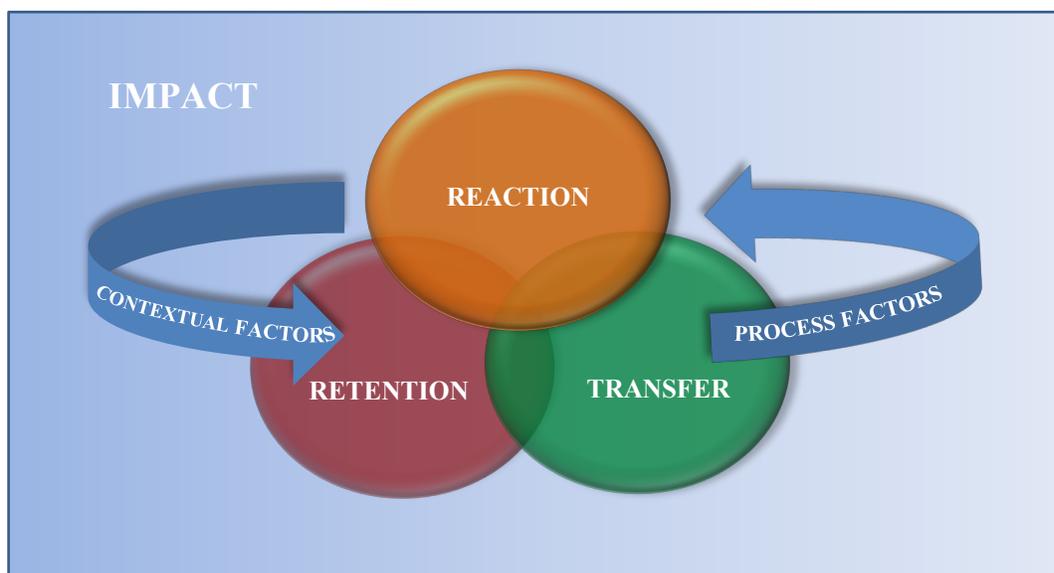


Figure 2. Proposed Evaluation Model

Research Design

Setting and Program Description

A developmental mathematics program at an urban research university, in the Midwest, will be evaluated using the proposed evaluation model. The program was founded in 2009 to

address the high failure rates in developmental math at the university. Patterned after the workshop model developed by Uri Treisman at the University of California at Berkeley, the program stands on the very simple but powerful belief – that all students have a unique and special greatness within them and that, through hard work and with the support of a caring community, that greatness can be realized. Two principles follow: (1) High standards and high expectations should be held at all times for every student in the program and (2) a sense of community should exist within the program - where every student is cared about, supported and made to feel that he/she belongs to something special. The program serves 450-500 students annually.

Participants

The participants for this study will be limited to a retrospective examination of participants of the workshop-model, developmental mathematics program and students who took the university's traditional developmental mathematics course. They will be grouped and tracked as cohorts by the semester in which they took the developmental course. Students who enrolled in the university's traditional developmental course served as the comparison group. Six semesters will be used for this study: Students who took the course during the terms of Fall 2012, Winter 2013, Fall 2013, Winter 2014, Fall 2014 and Winter 2015 will be placed into two groups. Students enrolled in the program's sections will be labeled Group A (N = 802) and students enrolled in the traditional sections will be labeled Group B (N = 2647).

Testing the New Model

The new model will be tested by using the following step-by-step process adapted from Booker (2016): The model was operationalized using the following steps: (1) identify key stakeholders, expectations and objectives, (2) identify program design and data sources, (3) collect

and analyze data, (4) define process and context factors of the program and link data points (5) provide recommendations.

The proposed model allows for phases to be revisited at any time during the evaluation while keeping institutional impact as the primary focus of the evaluation. Moreover, the interrelation of the phases will allow connections to be made between programmatic activities and outcomes.

Variables

The proposed evaluation model will be applied to the developmental mathematics program. The independent variables for this study include students in both developmental course options. Participants in Group A are students in the specially-designed developmental math program and received traditional face-to-face instruction along with a workshop. Participants in Group B are students who chose the university's traditional developmental math course. Instruction to students in Group B was delivered via online computer modules with two hours of face-to-face instruction, weekly. The dependent variables in the study will include final exam scores, course grades, grade point average (GPA) and graduation status.

Data Collection

This study will be retrospective-descriptive and causal-comparative (Johnson, 2001). All data collected will be retrospective. Data will be collected at the institutional, departmental and programmatic levels. Institutional level data will be collected from the Student Tracking and Retention System (STARS). Course grades, GPA and graduation status will be tracked through STARS. All information will be exported and maintained in a Microsoft Excel spreadsheet. Final exam data will be provided by the university's Mathematics Department. Results of the student satisfaction survey administered were provided at the end of each semester.

Data Analyses

The proposed evaluation model incorporates four stages. At Stage 1, reaction, the program's student satisfaction survey will be used to examine students' attitudes toward the program. The program issued the survey at the end of each of the semesters being investigated. Hard copies of the surveys are available. Frequencies will be used to analyze the results of the survey. At Stage 2, Retention, quiz and test scores will be analyzed. At Stage 3, Transfer, final exam scores will be used. An independent samples t-test will be conducted on final exam scores to determine if there is a statistical difference in student achievement between Group A and Group B. At Stage 4, Impact, final course grades will be used. Frequencies will also be used to analyze the distribution of course grades between the two groups. The Statistical Package for the Social Sciences (SPSS 23.0) was used to conduct the independent samples t-test with a significance level of $\alpha=0.05$. The contextual and process factors outlined above will be used to guide the analysis to detect causal linkages from program processes to outcomes.

Meta-evaluation

A meta-evaluation will be conducted to assess the efficacy of the proposed model and address the second research question, based on the standards from the Joint Committee on Standards for Education Evaluation. The five standards categories were summarized by Booker (2016):

1. Utility Standards which help to assure that stakeholders find program evaluation processes valuable to their needs (Joint Committee on Standards for Educational Evaluation, 1994).
2. Feasibility Standards which are intended to increase evaluation effectiveness and efficiency (Joint Committee on Standards for Educational Evaluation, 1994).

3. Propriety Standards which support legal, fair and just evaluations (Joint Committee on Standards for Educational Evaluation, 1994).
4. Accuracy Standards which are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality (Joint Committee on Standards for Educational Evaluation, 1994).
5. Evaluation Accountability Standards which encourage adequate documentation of evaluations and a meta-evaluative perspective focused on improvement and accountability for evaluation processes and products. (Joint Committee on Standards for Educational Evaluation, 1994)

The standards that follow under each category are as follows:

Utility Standards: U1) evaluator credibility, U2) attention to stakeholders, U3) negotiated purposes, U4) explicit values, U5) relevant information, U6) meaningful processes and products, U7) timely and appropriate communicating and reporting, and U8) concern for consequences and influence.

Feasibility Standards: F1) project management, F2) practical procedures, F3) contextual viability, and F4) resource use.

Propriety Standards: P1) responsive and inclusive orientation, P2) formal agreements, P3) human rights and respect, P4) clarity and fairness, P5) transparency and disclosure, P6) conflicts of interests, and P7) fiscal responsibility.

Accuracy Standards: A1) justified conclusions and decisions, A2) valid information, A3) reliable information, A4) explicit program and context descriptions, A5) information management, A6)

sound designs and analysis, A7) explicit evaluation reasoning, and A8) communication and reporting.

Evaluation Accountability Standards: E1) evaluation documentation, E2) internal meta-evaluation, E3) external meta-evaluation.

Stufflebeam's (2011) checklist was used to compare the 30 standards against the evaluation conducted using the proposed evaluation model. The checklist contains ratings for each of the 30 evaluation standards – with each standard being broken down into six checkpoints (Blake, 2013). The checklist scorecard by category is outlined in Table 3. As described by Blake (2013), the evaluator counts the number of checkmarks within each standard.

If the evaluator checked all six statements, the standard is assigned a rating of “Excellent.”

If the evaluator checked five of the six statements, the standard is assigned a rating of “Very

Good.” If four of the six statements are checked, the standard is assigned a rating of

“Good.” If two or three of the six statements are checked, the standard is assigned a rating

of “Fair.” If none of the statements are checked, the standard is assigned a rating of “Poor.”

(p. 75).

CHAPTER 4: RESULTS

Operationalization of Proposed Model

The model was operationalized using the following steps: (1) identify key stakeholders and objectives, (2) identify program design and data sources, (3) collect and analyze data, (4) define process and context factors of the program and link data points (5) provide recommendations.

Step 1: Identify key stakeholders and objectives. Key stakeholders for this this evaluation were identified as program participants, instructors, student support staff and program administrators. Due to the retrospective nature of the evaluation, stakeholder surveys, records and reports were used to examine through the program. Program participants' feedback was examined through a student satisfaction survey that was administered at the end of the semester in which they took the developmental mathematics course. Instructors and student support staff are responsible for maintaining class records which includes quiz and test scores, homework and attendance. Program administrator reports included passing rates and final exam data.

Step 2: Identify program design and data sources. The program used for this study was founded in 2009 to address the high failure rates in developmental math at the university. Patterned after the workshop model developed by Uri Treisman at the University of California at Berkeley, the program stands on the very simple but powerful belief – that all students have a unique and special greatness within them and that, through hard work and with the support of a caring community, that greatness can be realized. Two principles follow: (1) High standards and high expectations should be held at all times for every student in the program and (2) a sense of community should exist within the program - where every student is cared about, supported and made to feel that he/she belongs to something special.

Data sources at the programmatic level included administrative reports, class records and electronic gradebooks. Additionally, data sources at the departmental level included administrative reports and final exam reports. At the institutional level, the data source was the university's retention tracking system.

Step 3: Collect and analyze data at each stage of the proposed model. Data were collected and analyzed based on the four stages of the proposed evaluation model.

Stage 1: Reaction

At Stage 1, Reaction, the program's student satisfaction survey was used to examine students' attitudes toward the program. The survey uses a Likert scale with the ratings of "Strongly Agree," "Agree," "No Opinion," "Disagree," and "Strongly Disagree." The responses were coded on a scale of 0-5 with Strongly Disagree = 0 and Strongly Agree = 5. Student satisfaction results for the terms Fall 2012, Winter 2013, Fall 2013, Winter 2014, Fall 2014 and Winter 2015 are summarized in Table #1. Of the 802 students that enrolled, 456 students responded to the survey. For Question 1, of the survey, 92.7% of the 456 respondents strongly agreed or agreed that the program made them feel more confident in their ability to succeed in math. For Question 2, 78% strongly agreed or agreed that the program made math more enjoyable for them. For Question 3, 82.4% strongly agreed or agreed that the hard work required by the program was worthwhile. For Question 4, 90.1% of the respondents strongly agreed or agreed that they felt cared about as an individual in the program. For Question 5, 87.5% agreed or strongly agreed that they would recommend the program to other students.

Table 3.
Satisfaction Survey Results

Question	Strongly Agree Total (%)	Agree Total (%)	No Opinion Total (%)	Disagree Total (%)	Strongly Disagree Total (%)

1. PROGRAM X has made me more confident in my ability to succeed in math.	282(61.8%)	141(30.9%)	25(5.5%)	6(1.3%)	2(0.4%)
2. PROGRAM X has made math a more enjoyable subject for me.	209(45.8%)	147(32.2%)	61(13.4%)	29(6.4%)	10(2.2%)
3. The hard work required by PROGRAM X was worthwhile.	213(46.7%)	163(35.7%)	52(11.4%)	17(3.7%)	11(2.4%)
4. I felt cared about as an individual in PROGRAM X.	261(57.2%)	150(32.9%)	35(7.7%)	6(1.3%)	4(0.9%)
5. I would recommend PROGRAM X to other students.	299(65.6%)	100(21.9%)	36(7.9%)	14(3.1%)	7(1.5%)

A 5×5 singly ordered Kruskal-Wallis test was conducted on the data in Table 3, with the null hypothesis being the 5 rows are identically distributed. The result was statistically significant, $KW = 67.17$, $df = 4$, $p = 0.000$. This indicates the greater proportion of those “strongly agreeing” and “agreeing” was consistent for all five questions.

Stage 2: Retention

At, Stage 2, Retention, exam scores, attendance and homework completion were examined for the terms Fall 2012 (N=143), Winter 2013 (N=75), Fall 2013 (N=180), Winter 2014 (N=75) Fall 2014 (N=148) and Winter 2015 (N=109). Program participants were administered four exams each semester to formatively evaluate their learning. Exam averages for each term are reported in Figure 3. For the Fall 2012 term, participants had an exam average of 83.3%. The exam average for the Winter 2013 term was 83%. During the Fall 2013 term, the exam average was 82.6%. The exam average for Winter 2014 was 80% and 84.7% for the Fall 2014 term. The exam average for the Winter 2015 term was 82.7%. The Mann Kendall test for linear trend was not statistically significant spanning this Fall 2012 – Winter 2015 ($\tau = -.2$, $p = .57$). This indicates the average scores were stable during this period.

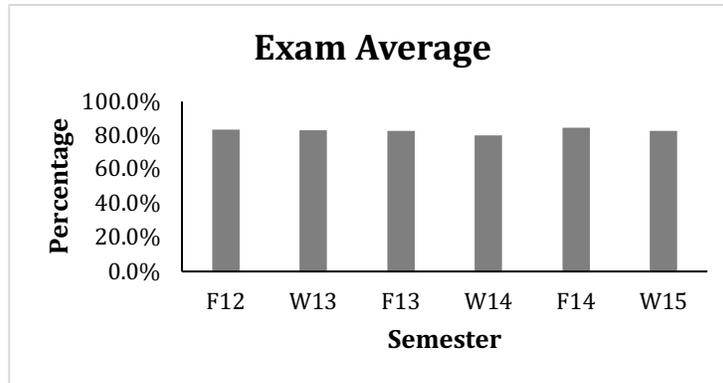


Figure 3. Exam Averages by Semester

Homework completion averages are presented in Figure 4. For the Fall 2012 and 2013 terms the average homework completion rate was 89.1%. The average homework completion rate for the Winter 2013 term was 90.9%. For the Winter 2014 term, the average was 81.8%. The averages for the Fall 2014 and Winter 2015 terms were 85.5% and 80%, respectively. Despite the visual decline from Fall, 2012 through Winter 2015, the Mann Kendall test for linear trend was not statistically significant ($\tau = .6$, $p = .14$), again indicated a stable pattern of homework completion.

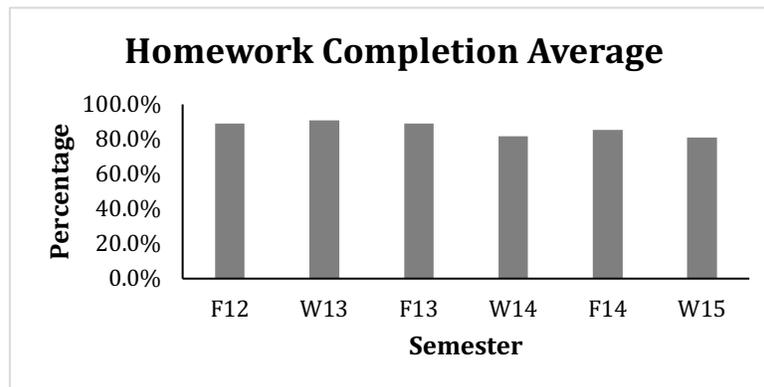


Figure 4. Homework Completion Averages by Semester

In Figure 5, the average class attendance is given. The average class attendance for the Fall 2012, Winter 2013 and Fall 2013 terms was 90.9%. The average class attendance for the Winter 2014 and Fall 2014 terms were 83.6% and 89.1%, respectively. The average class attendance for the Winter 2015 term was 83.6%. Despite the visual decline from Fall, 2012 through Winter 2015, the

Mann Kendall test for linear trend was not statistically significant ($\tau = .6$, $p = .14$), indicated a stable pattern of average class attendance.

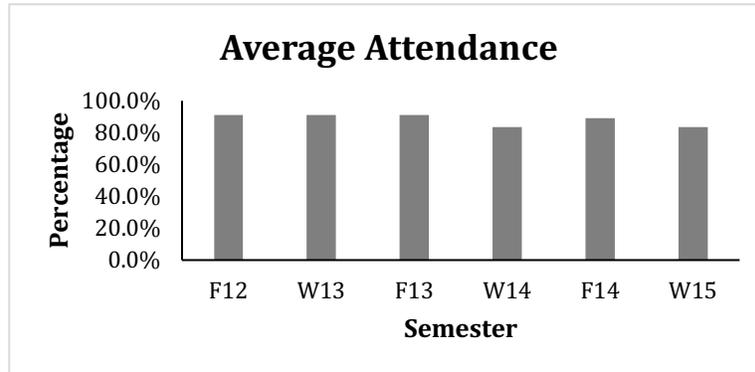


Figure 5. Average Class Attendance by Semester

Stage 3: Transfer

At Stage 3, Transfer, the scores from the departmental, cumulative final exam were analyzed for the Fall 2012, Winter 2013, Fall 2013, Winter 2014, Fall 2014 and Winter 2015 semesters. The program participants were administered the same final exam as students enrolled in regular sections of the developmental mathematics course. An independent samples t-test was conducted to see if the program participants' final exams scores differed from those of the students enrolled in the traditional sections. The program participants were labeled as Group A ($N=664$) and the students enrolled in the traditional sections were labeled as Group B ($N=1898$). The test was conducted with a nominal alpha level of $\alpha = 0.05$ and the following statistical hypotheses:

$$H_0 : \mu_{\text{GroupA}} = \mu_{\text{GroupB}}$$

$$H_1 : \mu_{\text{GroupA}} \neq \mu_{\text{GroupB}}$$

Shown in Table 4 are the mean final exam score for Group A, the program participants, was 81.24, and the mean final exam score was 71.42 for students enrolled in the traditional sections, Group B. Results from the t-test are presented found in Table 5. The independent samples t-test revealed a 2-tailed significance value of $p = 0.000$. Thus, the null hypothesis was rejected

and it was concluded that there was a statistically significant difference between the final exam means of Group A and Group B. The program participants in Group A scored statistically significantly higher on the final exam (81.24 ± 0.680) than the students in Group B who were enrolled in the regular sections (71.42 ± 0.405), $t(2560) = 12.353$, $p = 0.000$.

Table 4.
Group Statistics: Program Participants and Students Regularly Enrolled

	Group	N	Mean	Std. Deviation	Std. Error Mean
Scores	Group A	664	81.24	17.532	0.680
	Group B	1898	71.42	17.674	0.405

Table 5.
T-Test Table: Program Participants and Students Regularly Enrolled

		Equal Variances Assumed	Equal Variances Not Assumed
Levene's Test for Equal Variances	F	5.509	
	Sig.	0.019	
	t	12.353	12.401
	df	2560	1166.686
Test for Equality of Means	Sig. (2 tailed)	0.000	0.000
	Mean Difference	9.823	9.823
	Std. Error Difference	0.79524	0.79216
95% Confidence Interval of the Difference	Lower	8.264	8.269
	Upper	11.383	11.377

Stage 4: Impact

At Stage 4, Impact, final course grades between Group A and Group B were analyzed. The grading scale for developmental courses differed from standard grading. The grading scale translates to the following: Earning a grade of ANC means a student passed the course with an 90% average or above. NC stand for “no (degree) credit.” Moreover, BNC means a student passed the course with an average between 80-89% and CNC means a student passed the course with an average between 70-79%. A student failed the course and earned a grade UNC if they had an overall average or less than 70%. Students who failed to score 60% or higher on the final exam also earned a grade of UNC – regardless of overall average. A grade of W means a student withdrew from the course.

In Figure 5, grade distributions are given for Group A (N = 802) and Group B (N=2647). In Group A, 42.3% of the students earned a grade of ANC while only 10.3% of students in Group B earned an ANC. 28.4% of students in Group A earned a BNC and 24% of students in Group B earned a BNC. Group A had 10.2% of its students earn a CNC while 22.8% of students in Group B earned a CNC. Additionally, in Group A, 17% of the students earned a grade of UNC. Group B had 40.8% of the students earn a UNC. 2.1% of the students in both groups withdrew from the course. Group A had an overall passing rate of 80.1% while Group B had an overall passing rate of 57.1%.

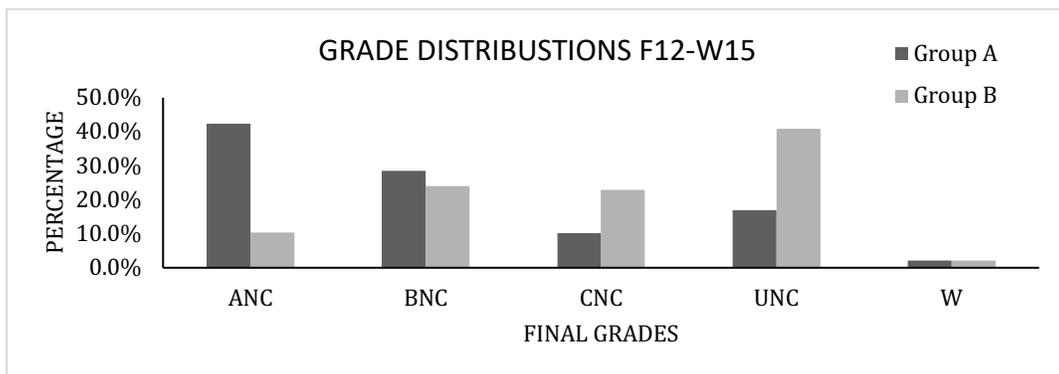


Figure 6. Grade Distributions: Group A vs Group B (Fall 2012-Winter 2015)

Step 4: Define process and context factors of the program. (1) history, (2) structure, and (3) policies. Additionally, there are four process factors defined within the proposed evaluation model: (1) collaboration, (2) community and (3) data collection.

Meta-evaluation

The meta-evaluation results of utility standards are shown in Table 6. The overall rating for utility was “Good” with a total score of 62.5%. The highest rating of “Excellent” was given to standard U6-Meaningful Processes and Purposes. U1-Evaluator Credibility and U4-Explicit Values were rated “Very Good” while all other standards were rated “Good.”

Table 6.
Utility Ratings

	Standard	Rating
U1	Evaluator Credibility	Very Good
U2	Attention to Stakeholders	Good
U3	Negotiated Purpose	Good
U4	Explicit Values	Very Good
U5	Relevant Information	Good
U6	Meaningful Processes & Purposes	Excellent
U7	Timeliness, Appropriate Communication and Reporting	Good
U8	Concern for Consequences & Influence	Good
Total Score: 62.5%		Overall Rating: Good

Shown in Table 7 are feasibility ratings. Each standard received a rating of “Very Good” for each of the four standards: F1-Project Management, F2-Practical Procedures, F3-Contextual Viability and F4-Resource Use. The overall feasibility rating was “Very Good” with a total score of 75%.

Table 7.
Feasibility Ratings

	Standard	Rating
F1	Project Management	Very Good
F2	Practical Procedures	Very Good
F3	Contextual Viability	Very Good
F4	Resource Use	Very Good
Total Score: 75%		Overall Rating: Very Good

Propriety ratings are shown in Table 8. The highest rating of “Excellent” was given to standards P3-Human Rights and Respect, P4-Clarity and Fairness and P5-Transparency and Disclosure. Standards P1-Responsiveness & Inclusive Orientation and P2-Formal Agreements were given a rating of “Very Good.” Standards P6 and P7 were given ratings of “Good” and “Fair,” respectively.

Table 8.
Propriety Ratings

	Standard	Rating
P1	Responsiveness & Inclusive Orientation	Very Good
P2	Formal Agreements	Very Good
P3	Human Rights and Respect	Excellent
P4	Clarity & Fairness	Excellent
P5	Transparency & Disclosure	Excellent
P6	Conflicts of Interest	Good
P7	Fiscal Responsibility	Fair
Total Score: 72.5%		Overall Rating: Very Good

Shown in Table 9 are accuracy ratings. The highest rating of “Excellent” was given to standards A2-Valid Information, A5-Information Management and A7-Explicit Evaluation Reasoning. A rating of “Very Good” was given to the standards A1-Justified Conclusions & Decisions, A3-Reliable Information, A4-Explicit Program & Context and A6-Sound Design and Analyses. Standard A8-Communication and Reporting received a rating of “Good.”

Table 9.
Accuracy Ratings

	Standard	Rating
A1	Justified Conclusions & Decisions	Very Good
A2	Valid Information	Excellent
A3	Reliable Information	Very Good
A4	Explicit Program & Context	Very Good
A5	Information Management	Excellent
A6	Sound Design & Analyses	Very Good
A7	Explicit Evaluation Reasoning	Excellent
A8	Communication & Reporting	Good
Total Score: 81.25%		Overall Rating: Very Good

Evaluation accountability ratings are given in Table 8. The overall rating was “Good” with a total score of 58.33%. The evaluation was rated “Very Good” on standards E1-Evaluation Documentation and E2-Internal Meta-evaluation. Standard E3-External evaluation was given a rating of “Poor”.

Table 10.
Evaluation Accountability Ratings

Standard		Rating
E1	Evaluation Documentation	Very Good
E2	Internal Meta-evaluation	Very Good
E3	External Meta-evaluation	Poor
Total Score: 50%		Overall Rating: Good

CHAPTER 5: DISCUSSION

The purpose of this study was to develop an evaluation model that can be used to evaluate educational programs using prospective or retrospective data. Kirkpatrick's (1959) framework was used to develop and propose a new evaluation model. Kirkpatrick's model has been criticized for its inability to connect programmatic outcomes to programmatic activities. The new model was able to address this limitation by providing causal linkages between programmatic activities/functions and programmatic outcomes by examining process and contextual factors. The new model was also able to identify key programmatic functions that were essential to replicating and/or scaling up successful programs.

The study also yielded a reimagined graphic of the model to better deliver its conceptual essence. Moreover, the stages are no longer numbered as numbers promote a linear approach to the evaluation process. The new design of the proposed model, shown in Figure 7, drives home its main idea that "impact" is at the core of the entire evaluation process. Each stage is linked to impact and the other stages simultaneously, while the contextual and process factors remain evaluated throughout the entire process. It is also important to note that each context and process factor can be prioritized and defined differently during different applications of the model.

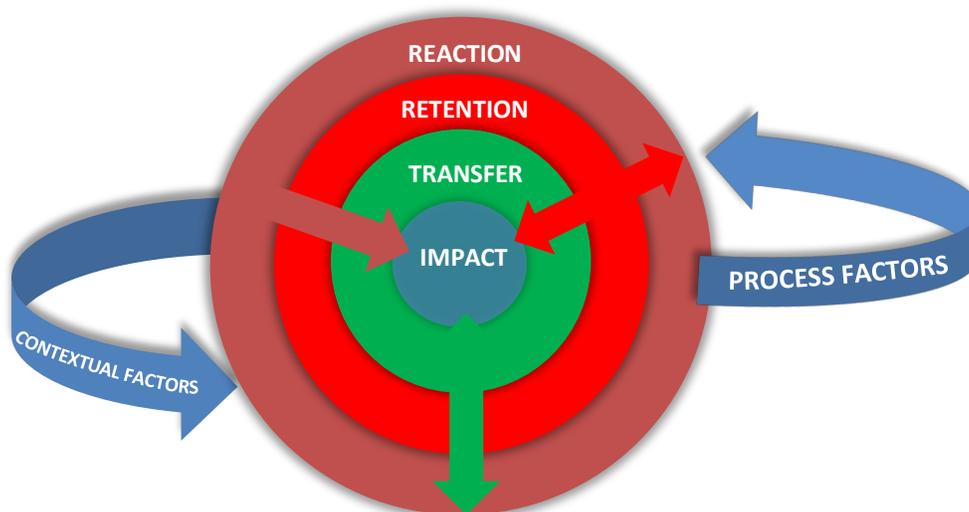


Figure 7. Proposed Model: Reimagined

This retrospective-descriptive and causal-comparative study was used to determine the effectiveness of the proposed evaluation model by addressing the following research questions:

- Research Question 1: Where does the new model fit in the body of evaluation research and theory?
- Research Question 2: How effective is the new model for detecting causal linkages between programmatic activities/functions and outcomes?
- Research Question 3: How well does the new model meet the Joint Committee's Program Evaluation Standard requirements for utility, feasibility, propriety, and accuracy?

Proposed Model Fit in Evaluation Research and Theory

The first research question was addressed during the review of the literature in Chapter 2. Despite the widespread use of Kirkpatrick's framework, it has been criticized for the lack of evidence that it provides in regard to causal linkages between training and outcomes. Kirkpatrick is vague about the precise nature of causal linkages between training activities and training outcomes, although he has made implications that simple causal linkages exists between the evaluation levels (Holton, 1996; Bates, 2004). Kirkpatrick (1994) has also asserted that it is important for trainees to react favorably, if training is going to be effective. Moreover, he stated that "without learning, no change in behavior will occur" (Kirkpatrick, 1994, pg. 51). However, research has largely failed to confirm such causal linkages (Bates, 2004). Meta-analyses of training evaluation studies using Kirkpatrick's framework have found little evidence either of substantial correlations between measures at different outcome levels or evidence of the linear causality suggested by Kirkpatrick (Bates, 2004). Thus, the framework is best described as a taxonomy rather than a model (Alliger & Janak, 1989). Moreover, Holton (1996) and Bates (2004) asserted

that causal conclusions are a necessary part of evaluation and a more complete model is required. The proposed model attempts to address these criticisms.

Classical evaluation models like Kirkpatrick and others mentioned throughout this study take a very traditional approach to evaluation – focusing on implementation and goal attainment. The proposed model encourages that impact is measured or examined at any or every stage of the process. Moreover, the proposed model calls for purposeful evaluation focusing on strategy, mission fulfillment, and organizational/systems change. The following principles are inherent in the proposed model.

1. All programmatic goals should be, in some way, connected to the organizational/institutional goals. Programmatic goals should always be written in alignment with the organization's strategic plan.
2. Organizational/institutional impact should be considered at each stage of the evaluation. Measuring impact is not a final stage – it is the goal throughout.
3. Any stage may be revisited throughout the evaluation process. The model does not require a linear application. Additionally, every stage may not be applicable - given the scope of the evaluation project.
4. All training/instructional goals should be clear, simple and measurable. Learners should have a clear understanding as to what they are expected to learn. Moreover, they should know how the information is prioritized.
5. A positive reaction to the learning experience does not necessarily mean learning or retention occurred. Similarly, a negative reaction to the learning experience does not necessarily mean that learning did *not* occur. Measuring reaction is an evaluation of the experience – not the instruction or training.

6. Measuring retention should not rely on memorization or recitation of facts. Retention should be measured on conceptual understanding, interpretation and application.
7. Demonstrated retention does not guarantee skills/knowledge will be transferable. Moreover, full transfer may not be measurable in educational settings.
8. Formative assessment of retention can lead to effective transfer. Formative assessment should also be used for continuous improvement of the instructional experience.
9. Instructor accountability is as important as learner accountability. Instruction should be evaluated independently of the learners – whenever possible.
10. Contextual and process factors should be used to determine causal linkages between programmatic activities and programmatic outcomes. Each factor can be prioritized differently – given the scope of the evaluation project.

Research questions (2) and (3) were addressed by operationalizing the proposed evaluation model. After a preliminary application of the proposed evaluation model, a meta-evaluation was conducted.

Key Findings

The process factors defined and applied within the model were (1) history, (2) structure and (3) policy. Additionally, the context factors defined and applied within the model were (1) collaboration and (2) community. These were used to summarize key findings about the program.

Process Factors

History: The developmental mathematics program was modeled after the university's honors mathematics program – which used the Treisman (1985) workshop model. The honors program, initiated in 1991, was specifically designed to increase minority and women participation in STEM. The following year, the program became one of the first in the country to use the

Treisman model at the pre-calculus level. By 1994, the program had begun to offer intermediate algebra courses using the Treisman (1985) model. After several years of success with minority students, the honors program was institutionalized and was given permanent funding in 1998. In 2008, due to high failure rates in the university's lowest developmental math course, the honors program's founder was approached to create the developmental math program – using the Treisman (1985) model. This proved to be significant due to the program's structure and policies. Participants have the same level of expectations as the honors students even though the course content is developmental.

Structure: The structure of the program's courses is very different from the traditional developmental courses. Students attend four, 50-minute, lectures per week and attend two, 85-minute, workshops. The workshop is a collaborative learning environment in which the students work in groups of three to four. For the first hour of the workshop, students complete sets of practice problems together. During the last 25 minutes, students work on very challenging problem that test their conceptual understanding. Collaboration is intentional with participants often collaborating outside of class. This has created a strong community of learners and an inclusive learning environment on campus.

Policies: The program has very high standards and expectations of participants. As a matter of policy, students are not allowed to enter lectures or workshops late - for any reason. Lectures and workshops start on-time and students are expected to be present at the start of each session. Moreover, failing to attend 80% or more of the lectures and/or completing 80% or more of homework assignments can result in failing the entire course. Homework is assigned at the end of each lecture. These policies led to students attending well over 80% of classes, on average with the average attendance - peaking over 90% for three of the six semesters of the term. Similarly,

the average homework completion rate was just as high during the six semesters examined. Over 82% of the respondents of the student survey strongly agreed or agreed that the hard work was worthwhile.

Context Factors

Collaboration: Program participants formally collaborate during the workshop component of the program. The workshop is designed for students to work in groups of three to four students. The groups change every workshop so that students are well-acquainted with every other student in the program. But again, this is intentionally designed to promote a sense of community and for collaboration to continue, informally, outside of the classroom. Which defined the second context factor, community.

Community: The program's community is made up of student participants, instructors, support coordinators and the director. Community building activities such as luncheons after each exam, mass exam reviews and game days are used to foster a sense of family amongst the student participants and staff. Student support coordinators are charged with supporting students with both academic and non-academic needs. Over 90% of the respondents of student survey strongly agreed or agreed that they felt cared about as an individual in the program with over 87.5% strongly agreeing or agreeing that they would recommend the program to other students. Also, 92.7% of the respondents strongly agreed or agreed that the program made them more confident in their ability to succeed at math.

Meta-Evaluation Findings

The meta-evaluation was conducted to address the third research question. According to Blake (2013), there is no generally accepted process for arriving at a final judgment of the usefulness, appropriateness, effectiveness, reliability, and validity of an evaluation model. Standards for

Program Evaluation (Yarbrough et al., 2011) is published by the Joint Committee on Standards for Educational Evaluation to help evaluators judge the quality of evaluations. Stufflebeam's (2011) checklist, which is aligned with the Joint Committee's standards, was used to see how well the proposed model and its application met the Joint Committee's Program Evaluation Standard requirements for utility, feasibility, propriety, and accuracy. Each category was rated on the scale of Poor, Fair, Good, Very Good and Excellent. It is also important to point out that this study is an evaluation of an evaluation design. Fitzpatrick et al. (2011) suggested that 18 of the 30 Joint Committee's Program Evaluation Standards (Yarbrough et al., 2011) are most relevant for evaluating evaluation designs. The standards and the respective ratings for this study are summarized in Table 11.

Table 11.
Relevant Meta-evaluation Standards for Evaluation Designs

Meta-evaluation standard	Rating
A6 – Sound Design and Analyses	Very Good
U3 – Negotiated Purpose	Good
P1 – Responsiveness and Inclusion Orientation	Very Good
P3 – Human Rights and Respect	Excellent
P4 – Clarity and Fairness	Excellent
P5 – Conflicts of Interest	Excellent
A1 – Justified Conclusions and Decisions	Very Good
A2 – Valid Information	Excellent
A3 – Reliable Information	Very Good
A4 – Explicit Program and Context Descriptions	Very Good
A5 – Explicit Evaluation Reasoning	Excellent
U2 – Attention to Stakeholders	Good
U4 – Explicit Values	Very Good
U6 – Meaningful Processes and Purposes	Excellent
F1 – Project Management	Very Good
F2 – Practical Procedures	Very Good
F3 – Contextual Viability	Very Good
F4 – Resource Use	Very Good

Utility Standards help to assure that stakeholders find program evaluation processes valuable to their needs (Joint Committee on Standards for Educational Evaluation, 1994; Booker, 2016). The model's application was rated overall "Good" for utility. According to Fitzpatrick et al. (2011), there are four utility standards that are relevant to evaluating an evaluation model/design: U2 – Attention to Stakeholders, U3 – Negotiated Purpose, U4 – Explicit Values, and U6 – Meaningful Processes and Purposes. The meta-evaluation yielded a rating of "Good" for standards U2 and U3. This was due to the informal nature of the evaluation which was conducted at no cost to the program. The evaluation was stronger for the standards U4 – Explicit Values and U6 – Meaningful Processes and Purposes. For U4, the model was rated "Very Good". Additionally, for U6, the evaluation was very strong – hitting every checkbox on Stufflebeam's (2011) checklist – receiving a rating of "Excellent".

Feasibility Standards are intended to increase evaluation effectiveness and efficiency (Joint Committee on Standards for Educational Evaluation, 1994; Booker, 2016). The meta-evaluation yielded an overall feasibility rating of "Very Good". All four feasibility standards, F1 – Project Management, F2 – Practical Procedures, F3 – Contextual Viability and F4 – Resource Use, are considered relevant by Fitzpatrick et al. (2011) when evaluating evaluation designs. Each standard received a rating of "Very Good" and an overall score of 75%.

Four of the seven propriety standards are relevant to evaluating evaluation designs (Fitzpatrick et al., 2011). They are: P1 – Responsiveness and Inclusion Orientation, P3 – Human Rights and Respect, P4 – Clarity and Fairness and P5 – Conflicts of Interest. Proprietary Standards support legal, fair and just evaluations (Joint Committee on Standards for Educational Evaluation, 1994; Booker, 2016). The study gave careful considerations to the propriety standards and it

received an overall rating of “Very Good” with an overall score of 78.5%. The evaluation was particularly strong in the areas of clarity, fairness and transparency.

The accuracy standards are intended to ensure that an evaluation uses sound theory and reasoning, as well as a solid design, in order to minimize inconsistencies and misconceptions and to produce thoughtful and truthful evaluation findings and conclusions (Stufflebeam, 2011; Blake, 2013). The highest score of the meta-evaluation was given to accuracy with a total score of 81.25% and overall rating of “Very Good”. Standards A2 – Valid Information and A5 – Information Management received ratings of “Excellent”. Both are considered to be relevant to evaluating evaluation designs by Fitzpatrick et al. (2011). Standards A1 – Justified Conclusions and Decisions, A3 – Reliable Information, A4 – Explicit Program and Context Descriptions, and A6 – Sound Design and Analyses are also given this consideration and the evaluation scored a rating of “Very Good” for each of those standards.

The final set of standards call for evaluator accountability. The evaluation received an overall rating of “Good” with an overall score of 50%. Although 50% is considered good for this standard, the lack of an external meta-evaluation creates a potential limitation to this study due to bias.

Limitations

Generalizability or external validity is a limitation to this study. Participants were from an urban research institution of higher education in the Midwest. The results may not translate to students taking developmental mathematics at other institutions. Moreover, the program had higher standards and requirements than the traditional developmental sections. Students self-selected into the program in which they knew required more work. Therefore, another limitation

are the threats to internal validity as it relates to selection and maturation. Lastly, as previously stated, the lack of an external meta-evaluation also limits the results of this study.

Recommendations for Future Research

The goal of this research was to offer an alternative evaluation model to Kirkpatrick's framework and to examine the effectiveness of the new method. More studies should be conducted using the proposed model to offer further critique and validation. Future research could also be done to build on the new framework to examine which process and/or context factors provide the strongest evidence of causal relationships between programmatic activities and outcomes. Additionally, retrospective data was used to complete the evaluation in this study. A study of the proposed model using prospective or a combination of prospective and retrospective data could also highlight other potential strengths and weaknesses of the model.

Moreover, more research should be conducted to determine ways to improve other existing frameworks and methodologies. This will not only strengthen the field of evaluation research but will ultimately strengthen programs, organizations and institutions that are subject to review and evaluation.

Implications for Practice

A principal goal of this study is to contribute to the research and field of evaluation. Although the field continues to develop, it is still dominated by classical models that are not intentional or focused on the why's behind the what's. Evaluation is essential to the decision making at all levels of an organization/institution and they should provide some contextual understanding of causal linkages. Although the proposed model is heavily influenced by Kirkpatrick's framework, it introduces a method for evaluators to connect programmatic activities with outcomes. Moreover, unlike Kirkpatrick, the proposed model encourages users not to use a

linear approach when analyzing information/data at the given stages. This offers opportunity for more robust evaluations and analyses and more emphasis on drawing connections to institutional impact.

The proposed model should be considered as one of many well-established evaluation tools to add to an evaluator's toolkit. Models offer the necessary structure to keep evaluation activities organized and systematic, however, evaluation requires flexibility and adaptability and one approach or methodology is rarely sufficient. Effective evaluation studies often use mixed methodologies and a combination of frameworks. Moreover, large scale evaluations can be time consuming and costly. Because of this, many of the classic models have been difficult to completely implement. A phased approach that connects each phase to organizational impact is key in establishing causal linkages and can potentially save time and money.

Implications Beyond Evaluation and Research Theory

Drawing on Wiggins and McTighe's (2011) instructional theory of backward design, the proposed model can be "flipped" and used as a learning design and developmental tool. The three stages of backward design are (1) Plan with the end in mind by first clarifying the learning you seek; (2) Think about the assessment evidence needed to show that students have achieved that desired learning; (3) Plan the means to the end – the teaching and learning activities and resources to help them achieve goals (Wiggins & McTighe, 2011). The proposed model can be operationalized as a learning design tool by using the following steps as adapted from Thalheimer (2018).

1. Consider all possible results. Examine all impacts the program might have on a wide range of stakeholders. Examine both positive and negative effects. Also, examine the potential effects the learning program may have on the institutional environment.

2. Examine how transfer or performance will be manifested. What transfer results are hoped to be achieved? Consider if full transfer is necessary and how it will be measured.
3. Create a comprehensive list of evaluation objectives outlined with measurable success indicators. Develop targeted measures for the retention and reaction stages.
4. After negotiating acceptability of evaluation objectives from key stakeholders, begin the instructional design process. Determine the targeted performance situations and targeted performance objectives in those situations.
5. Begin developing and designing the learning program. Be sure to get input from key stakeholders before implementation.
6. Deploy the learning program and evaluate its effectiveness at the various stages. Make necessary changes to improve the instructional process as it is being delivered.
7. Gather and analyze data and report/share it with key stakeholders. Request formal feedback from stakeholders.
8. Evaluate routinely for continuous improvement. Connect evaluation to program and institutional goals – whenever possible.

Summary and Conclusion

There are many different models, frameworks and methodologies in the field of program evaluation. Virtually all methodologies have merit and have made a substantial contribution to the field. In addition to examining effectiveness, program evaluation should offer more than what the outcomes are – particularly for programs that have been deemed successful over a sustained amount of time. This offers the best direction in regard to expansion, replication and/or sustainability. Failure to establish causal relationships can lead to “boutique” programming that offers little impact to the institution as a whole.

This study proposed an evaluation model that was influenced by Kirkpatrick's (1959) original framework. It consisted of the proposed model being used to conduct an evaluation of a developmental mathematics program. The evaluation and its design were subsequently evaluated through a thorough meta-evaluation process guided by the Joint Committee Standards for Program Evaluation (Yarbrough et al., 2011) and Stufflebeam's (2011) meta-evaluation checklist. The evaluation and its design yielded favorable ratings from the meta-evaluation – preliminarily validating its merit and fit in evaluation theory and research.

The proposed model places emphasis on drawing connections between programmatic activities and programmatic outcomes. The Kirkpatrick framework was improved by defining process and context factors to consider throughout the evaluation process. Such considerations allowed for conclusions to be drawn as to how those factors influenced the program's outcomes. Moreover, according to Pulichino (2007) and Kennedy (2012), the Results phase of the Kirkpatrick model was rarely used and thus, impact was rarely examined during its application. The proposed model calls for impact to be examined at each phase and allows the evaluation to move back and forth between phases. Future research will be conducted to further establish the new model.

REFERENCES

- American Society for Training & Development. (2009b). The value of evaluation: Making training evaluations more effective. Alexandria, VA: ASTD Research.
- Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42*, 331-342.
- Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning, 27*(3). 341–347.
- Booker, Z. (2016). A new logic model for change. Doctoral dissertation. Wayne State University, Detroit, MI.
- Christie, C. A., & Alkin, M. C. (2013). An evaluation theory tree. Evaluation Roots: A Wider Perspective of Theorists' Views and Influences, 11-57.
- Cole, T. (2015). *Curriculum Evaluation: Goal Based vs Goal Free*. Retrieved from <https://theelasticscholastic.wordpress.com/2015/02/28/curriculum-evaluation/>
- Cooksy, L. J. & Caracelli, V. J. (2009). Meta-evaluation in practice: Selection and application of criteria. *Journal of Multidisciplinary Evaluation, 6* (11). 1-15.
- Dubin, R. (1976). Theory building in applied areas. In *M. D. Dunnette (Ed.), Handbook of Industrial/Organizational Psychology* (p. 17–39). Skokie, IL: Rand McNally
- Dunin-Woyseth, H. & Nillson, F. (2013) Doctorateness in design disciplines. Negotiating connoisseurship and criticism in practice-related fields. *FORMakademisk 5*(2). 1-11.
- Eisner, E.W. (1979). *The educational imagination: On the design and evaluation of school programs*. New York: Macmillan.
- Fitzpatrick, J., Sanders, J., & Worthen, B. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th Ed.). New York: Allyn & Bacon.

- Fox, R.D. (2011) Revisiting “Discrepancy analysis in continuing medical education: A conceptual model.” *Journal of Continuing Education in the Health Professions*, 31(1). 71-76.
- Greene, J.G. (1997). Participatory evaluation. In L. Mabry (Ed.), *Advances in program evaluation: Evaluation and the post-modern dilemma* (p. 171-189) Greenwich, CT: JAI Press.
- Guba, E., & Lincoln, Y. (1981). *Effective evaluation*. San Francisco: Jossey-Bass.
- Gunderman, R. B., & Chan, S. (2015). Kirkpatrick's Evaluation of Educational Programs and its Relevance to Academic Radiology. *Academic Radiology*, 22(10). 1323-1325.
- Hill, J. C. (1986). *Curriculum evaluation for school improvement*. Springfield, Ill., USA: C.C. Thomas.
- Holton, E. F., III. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7(1). 5-29.
- Hogan, R.L. (2007). The historical development of program evaluation: exploring the past and present. *Online Journal of Workforce Education Development*, 2(6). 1-14.
- Johnson, B. (2001). Toward a new classification of nonexperimental quantitative research. *Educational Researcher*, 30(2), 3-13.
- Joint Committee on Standards for Educational Evaluation (1994). *The Program evaluation standards* (2nd ed.). Thousand Oaks, California: Sage Publication.
- Kennedy, P. E. (2012). Training professionals’ usage and understanding of Kirkpatrick’s level 3 and level 4 evaluations. Master’s thesis. Boise State University, Boise, ID.
- Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs: Reaction. *American Society for Training and Development Journal*, 18. 3-9.
- Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs: Learning. *American Society for Training and Development Journal*, 18. 21-26.

- Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs: Behavior. *American Society for Training and Development Journal*, 19, 13-18.
- Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs: Learning. *American Society for Training and Development Journal*, 18, 28-32.
- Kirkpatrick, D.L. (1994). *Evaluating Training Programs*. San Francisco: Berrett-Koehler Publishers, Inc.
- Klimoski, R. (1991). Theory presentation in human resource management. *Human Resource Management Review*, 4, 253–271.
- Madaus, G. F., & Stufflebeam, D. L. (1988). Educational evaluation: The classical writings of Ralph W. Tyler. Boston: Kluwer
- McKenna, C. (1981) Making evaluation meaningful. *Journal of Extension*, 19(5), 9-14.
- Mertens, D. & Wilson, A. (2012). *Program evaluation theory and practice: A comprehensive guide*. New York: Guilford Press. EI
- Nilsen P (2015). Making sense of implementation theories, models and frameworks. *Implement Sci.* 10, 53.
- Patton, Q. M. (1997). *Utilization focused evaluation: The new century text (3rd Ed.)*, London: Sage Publications.
- Provus, M. (1971). Discrepancy Evaluation. Berkeley, CA: McCutchan.
- Pulichino, J. P. (2007). Usage and value of Kirkpatrick's four levels of training evaluation. (Unpublished doctoral dissertation). Pepperdine University, Malibu, CA.
- Reio, T., Rocco, T., Smith, D. & Chang, E.. (2017). A Critique of Kirkpatrick's Evaluation Model. *New Horizons in Adult Education and Human Resource Development*, 29, 35-53.
- Scriven, M. (1972). Pros and cons about goal-free evaluation. *Evaluation Comment*, 3(4), 1–4.

- Smith, M. K. (2005). *Elliot W. Eisner, connoisseurship, criticism and the art of education* The encyclopedia of informal education.
- Stake, R. E. (1975). *Evaluating the Arts in Education: A Responsive Approach*. Columbus, OH: Merrill.
- Stufflebeam, D. L. (2001). Evaluation models. *New Directions for Evaluation*, 89, 7-98.
- Stufflebeam, D. (2003). The CIPP model of evaluation. *Springer international handbooks of education: International handbook of educational evaluation*.
- Smith, E.R. & Tyler, R.W. (1942). *Appraising and recording student progress*. New York: Harper & Row.
- Taylor, P. & Cowley, D. (1972). *Readings in curriculum evaluation*. Dubuque, Iowa: W.C. Brown Co.
- Thalheimer, W. (2018). *The learning-transfer evaluation model: Sending messages to enable learning effectiveness*. Retrieved from <https://WorkLearning.com/Catalog>
- Thanabalan, V.T., Siraj, S. & Alias, N. (2015). Evaluation of a digital story pedagogical module for the indigenous learners using the stake countenance model. *Turkish Online Journal of Educational Technology*, 14(2). 63-72.
- Treisman, P. U. (1985). A study of the mathematics performance of Black students at the University of California, Berkeley. Unpublished doctoral dissertation, University of California, Berkeley.
- Tyler, R.W. (1942). General statement on evaluation. *Journal of Educational Research*, 35(7). 492-501.
- Warju, W. (2016). Educational program evaluation using CIPP model. *Innovation of Vocational Technology Education*, 12(1). 36-42.

- Watkins, W., Leigh, D., Foshay, R. & Kaufman, R. (1998) Kirkpatrick Plus: Evaluation and Continuous Improvement with a Community Focus. *Educational Technology Research and Development*, 46(4). 90-96.
- Wiggins, G., and McTighe, J. (2011). The Understanding by Design guide to creating high-quality units. Alexandria, VA: ASCD.
- Worthen, B. R. and Sanders, J. R. (1987). *Educational Evaluation: Alternative Approaches and Practical Guidelines*. New York: Longman.
- Worthen, B.R., Sanders, J.R., & Fitzpatrick, J.L. (2004). *Educational evaluation: Alternative approaches and practical guidelines*. (3rd ed.). Boston: Allyn & Bacon.
- Youker, B.W. & Ingraham, A. (2013). Goal-free evaluation: An orientation for foundations' evaluations. *Foundation Review* 5(4). 51-61.
- Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Williams, J., Shea, C. & Misulis, K. (2011). Using the context, input, process, and product evaluation model (CIPP) as a comprehensive framework to guide the planning, implementation, and assessment of service-learning programs. *Journal of Higher Education Outreach and Engagement* 15(4). 57-84.

ABSTRACT**BEYOND THE FOUR LEVELS: AN EVALUATION MODEL FOR GROWTH AND SUSTAINABILITY**

by

DARRYL GARDNER**May 2019****Advisor:** Dr. Shlomo Sawilowsky**Major:** Education Evaluation and Research**Degree:** Doctor of Philosophy

Given the limitations in the Kirkpatrick evaluation model, the purpose of this study was to develop an evaluation model that can be used to evaluate educational programs using prospective or retrospective data. Kirkpatrick's original framework was used to develop a new evaluation model that took into account contextual and process factors. To examine the efficacy of the new model, a meta-evaluation was conducted using Stufflebeam's (2011) Program Evaluations Checklist. The checklist is based on the Joint Committee on Standards for Educational Evaluation (2011). The new model received favorable scores in the categories of utility, feasibility, propriety, accuracy and evaluator accountability. It was concluded that new model was able to address the major flaw of the Kirkpatrick's model by providing causal linkages between programmatic activities/functions and programmatic outcomes.

AUTOBIOGRAPHICAL STATEMENT

Darryl Gardner is native of Detroit and a product of the Detroit Public School system. He holds a Ph.D. in Evaluation and Research (Applied Statistics) from Wayne State University. His education also includes a B.A. degree in Mathematics and an M.Ed. in Learning Design and Technology – also from WSU.

With nearly 15 years of experience in higher education, Darryl currently serves as director of the Emerging & Rising Scholars Programs within the Center for Excellence & Equity in Mathematics at WSU. He also serves as the dean of students for the nationally-recognized Math Corps program and has led the program's national expansion efforts in Philadelphia, PA. Moreover, Darryl is extremely active in his community. He spends a significant amount of time advocating for foster youth and promoting mathematics literacy.

Darryl is a proud member of Phi Beta Sigma Fraternity, Inc. His hobbies include playing piano, chess and basketball with his son, Cameron.