EXPLORING THE LEVEL OF EVIDENCE OF IMPACT COLLECTED THROUGH PROGRAM EVALUATION BY COOPERATIVE EXTENSION AS PUBLISHED IN THE JOURNAL OF EXTENSION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Jeffrey Dale Workman, B.S., M.S.
Graduate Program in Agricultural and Extension Education

The Ohio State University

2010

Dissertation Committee:

Dr. Scott D. Scheer, Advisor

Dr. Joseph A. Gliem

Dr. Karen S. Bruns
Documenting evidence of impact demonstrates that organizations and programs are making a difference and/or accomplishing something positive. The results of a program must ultimately change people’s attitudes or behavior, or benefit society in other ways (Diem, 1997). As competition increases and the availability of financial resources decreases, the U.S. Cooperative Extension System must document evidence of impact to demonstrate that it’s worth funding.

In general, Extension programs which utilize program evaluation in the form of outcome studies do not document impact by collecting real evidence of behavior change and/or greater end results that benefit societal conditions through social, economic, civic, or environmental improvement. Traditionally, the criticism of evaluation in Extension has been that evaluators have simply reported on the number and demographics of the people who participated (Gentry-Van Laanen & Nies, 1995).

The purpose of this study was to explore and describe the level of evidence of impact collected through program evaluation (outcome studies) by the Cooperative Extension System as published in the *Journal of Extension* (JOE); and to examine any potential relationships between the documentation of evidence of impact with reach (size of audience), richness (detail and depth of interaction), and geographic scope. Articles reviewed in JOE were those listed under the headings of ‘Feature Articles’ and ‘Research
in Brief" during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. This study utilized a form of quantitative content analysis where data was collected and coded using qualitative methods and analyzed via quantitative techniques.

Two evaluation tools were used for this study. Bennett’s Hierarchy was utilized because it was considered to be the simplest, easiest to understand, and contained the most applicable description of different levels of evidence of impact. The University of Wisconsin-Extension logic model format was also included because it was considered to be the more contemporary and predominant tool currently used when documenting and describing results from Extension programming.

There were a total of 1,246 articles investigated in which 302 were selected for data collection. This investigation clearly showed that the number of evaluative studies with outcome data has increased going from 16 articles in the 1960s to 150 articles in the 2000s. The findings indicated that 88.5% of the articles documented evidence above the level of participation, almost two-thirds were measuring outcomes, about one-third were measuring evidence of substantial impact, and 5.6% documented “real or actual” impact. The relationships of reach of richness with level of evidence were low at best, and the relationship with geographic scope was very low.

It is highly likely that Extension programs make a tremendous positive difference in the lives of individuals and their communities and Extension greatly benefits society as a whole, but more program evaluation with evidence of higher-level impact is needed to support this argument. It is also likely that funding becomes more limited and much more competitive in the future. Those organizations who have real documented evidence
of their worth and merit will be the ones that receive the funding and continue to exist in the future.
DEDICATION

To my parents:
Robert & Ladonna Workman

To my grandparents:
David & Zelda Workman

Alfred & Maryland Henderson (In memoriam)
ACKNOWLEDGMENTS

I am very grateful for the help and support of family, friends, colleagues, instructors, and mentors in pursuing this academic endeavor and completing this dissertation. I would like to express my sincere gratitude to Dr. Scott Scheer for his many years of service as my academic advisor. I especially enjoyed and learned a great deal from our spirited discussions during our many meetings. I truly believe that our discussions helped to make the research and contents of this document much better. I would also like to express my appreciation to Drs. Joe Gliem and Karen Bruns for their contribution and agreeing to serve on my dissertation committee. I would like to thank Dr. Ryan Schmiesing for all of his help and suggestions. It was a tremendous help to bounce ideas and thoughts off of someone else and receive an additional professional perspective. I appreciate Dr. Ian Wilkinson’s assistance in serving as the Graduate School Representative during the dissertation defense.

I wish to acknowledge my candidacy committee which consisted of Drs. Robert Birkenholz and Dorinda Gallant in addition to Drs. Scheer and Gliem. I would also like to thank the various Specialists from Veterinary Extension; most especially Drs. Gustavo Schuenemann, William Shulaw, Kent Hoblet, William Saville, and Gary Bowman for allowing me to gain background experience and providing inspiration for this work.
VITA

November 27, 1973 ........................................... Born – Bellaire, Ohio

1992 .............................................................. High School Diploma, Bellaire High School

1997 .............................................................. B.S. in Agriculture, The Ohio State University

                                   Major: Animal, Dairy, and Poultry Science
                                   Minor: Agribusiness

2000 .............................................................. M.S., The Ohio State University

                                   Major: Agricultural Education
                                   Specialization: Extension Education

2000 .............................................................. Graduate Research Associate, Department of Human and Community Resource Development, The Ohio State University

2003 – Current ............................................ Extension Program Assistant, Veterinary Extension, Department of Veterinary Preventive Medicine, The Ohio State University

vii
PUBLICATIONS

Refereed Journal Articles


Refereed Papers & Presentations


Schuenemann, G. M., Bas, S., Workman, J. D., & Rajala-Schultz, P. (2009, September). *Beyond perception: Assessing reproductive modules within a continuing education program for dairy veterinarians*. Poster session presented at the 42nd Annual Conference of the American Association of Bovine Practitioners, Omaha, NE.


Booklets


**Magazine Articles**


**Newsletter Articles**


**FIELDS OF STUDY**

Major Field: Agricultural and Extension Education

Specialization: Extension Education

Minor Fields: Program Evaluation

Research and Statistics
# TABLE OF CONTENTS

Abstract ........................................................................................................................................ ii

Dedication ................................................................................................................................... v

Acknowledgments....................................................................................................................... vi

Vita........................................................................................................................................ vii

List of Tables .............................................................................................................................. xiv

List of Figures ............................................................................................................................. xvi

Chapter 1: Introduction .............................................................................................................. 1

  Background – Cooperative Extension .................................................................................... 1

  Background – Program Evaluation .......................................................................................... 7

  Background – Impact ............................................................................................................. 12

  Background – Evaluation Tools for Documenting Evidence of Impact ............................ 13

  Background – Reach, Richness, & Geographic Scope ......................................................... 16

Statement of Problem .............................................................................................................. 19

Purpose of the Study .................................................................................................................. 21

Objectives of the Study ............................................................................................................. 21

Conceptual Framework ............................................................................................................ 22
Data Collection.............................................................................................................. 77

Data Analysis ................................................................................................................ 81

Chapter 4: Results............................................................................................................. 84

Objective 1 .................................................................................................................... 87

Objective 2 .................................................................................................................... 89

Objective 3 .................................................................................................................... 91

Objective 4 .................................................................................................................... 96

Objective 5 .................................................................................................................... 99

Objective 6 .................................................................................................................. 100

Objective 7 .................................................................................................................. 101

Chapter 5: Summary, Discussion, & Conclusions .......................................................... 103

Summary ..................................................................................................................... 103

Summary of Limitations .............................................................................................. 105

Summary and Discussion of Findings .......................................................................... 106

Objective 1 .................................................................................................................... 107

Objective 2 .................................................................................................................... 109

Objective 3 .................................................................................................................... 113

Objective 4 .................................................................................................................... 116

Objective 5 .................................................................................................................... 118
LIST OF TABLES

Table 1: Stage I Cross Tabulation between Researcher and Rater A ....................... 73
Table 2: Stage I Cross Tabulation between Researcher and Rater B ....................... 74
Table 3: Stage II Percent Agreement between Researcher and Raters by Variable .... 75
Table 4: Stage I Cross Tabulation between Researcher’s First and Second Selections .. 76
Table 5: Stage II Percent Agreement between Researcher’s First and Second .......... 77
Table 6: Occupations of Lead Authors for Articles Selected during Stage I .......... 86
Table 7: Frequencies, Percentages, & Mean Rank for Objects of the Evaluation ...... 87
Table 8: Articles Reviewed for this Study ................................................................. 88
Table 9: Program Areas Represented in Selected Articles .................................... 89
Table 10: Highest Level of Evidence Obtained for Bennett’s Hierarchy .................. 90
Table 11: Highest Level of Evidence Obtained for UWEX Logic Model ................. 90
Table 12: Percentile of articles According to Level of Evidence ......................... 91
Table 13: Bennett’s Hierarchy by Decade ............................................................. 92
Table 14: UWEX Logic Model by Decade ............................................................ 93
Table 15: Bennett’s Hierarchy Groupings by Decade ........................................... 94
Table 16: UWEX Logic Model Groupings by Decade .......................................... 94
Table 17: F & % of Articles Documenting Evidence of Substantial Impact ............ 95
Table 18: Frequencies, Percentages, & Mean Rank for Decade ......................... 96
LIST OF FIGURES

Figure 1: Conceptual Framework ..................................................................................... 23
Figure 2: Model for Extension Education Approaches..................................................... 37
Figure 3: Kirkpatrick’s Four-Level Model ....................................................................... 39
Figure 4: A Hierarchy of Evidence for Program Evaluation............................................ 41
Figure 5: Bennett’s Hierarchy-Contemporary Example ................................................... 42
Figure 6: Program Action – Logic Model......................................................................... 45
Figure 7: Logic Model and Common Types of Evaluation .............................................. 47
Figure 8: Traditional Economics of Information.............................................................. 54
Figure 9: Articles Selected as Program Evaluations with Outcome Data ....................... 88
Figure 10: Bennett’s Hierarchy by Decade..................................................................... 92
Figure 11: UWEX Logic Model by Decade..................................................................... 93
Figure 12: Trend of Mean Ranks...................................................................................... 96
CHAPTER 1

INTRODUCTION

Documenting evidence of impact demonstrates that organizations and programs are making a difference and/or accomplishing something positive. As competition increases and the availability of financial resources decreases, the U.S. Cooperative Extension System must demonstrate that it’s worth funding. Cooperative Extension must create new programs and improve existing programs so that programming continues to make an impact and positively benefits society. This chapter will introduce the Cooperative Extension System, program evaluation, impact, evaluation tools, and reach/richness/geographic scope. In addition, this chapter will identify the problem, purpose, objectives, limitations, assumptions, and significance of this study.

Background – Cooperative Extension

The Cooperative Extension System (also referred to as CES, Cooperation Extension, or Extension) is the largest adult education system in the nation (Franz & Townson, 2008). Each U.S. state and territory has a state office at its land-grant university and a network of local or regional offices which are staffed by one or more experts who provide useful, practical, and research-based information to agricultural producers, small business owners, youth, consumers, and others in rural areas and
communities of all sizes (USDA NIFA, 2010). The foundation of Cooperative Extension in the United States began in 1862 with the Morrill Act establishing land-grant colleges, the Organic Act establishing the United States Department of Agriculture (USDA), and the Homestead Act providing settlers the opportunity to claim land. Claimants were required to “improve” the 150 acre plot by building a dwelling and cultivating the land for five years, and they would receive the tract free and clear (Homestead Act of 1862). The population who received this new land needed information on how to farm (Cross, 1999). In 1887, the Hatch Act created a connection between the USDA and the land-grant colleges by providing funding for experiment stations where researchers could create new knowledge. In addition, the Hatch Act of 1887 also required the mailing of bulletins, reports, periodicals, reprints of articles, and other publications necessary for the dissemination of results of the research and experiments. The land-grant colleges initially attempted to do this through written publications and Farmer’s Institutes; however, a much better system was needed.

The original Agricultural Extension Service was formally established by the Smith-Lever Act of 1914 “in order to aid in diffusing among the people of the United States useful and practical information on subjects relating to agriculture and home economics and to encourage the application of the same” (Rasmussen, 1989). Since that time, the name has changed to the Cooperative Extension System to better signify the cooperation and funding partnership between the land-grant universities, the USDA, and the local communities (county governments). Numerous legislative acts have been
passed since its inception, which have shifted the focus and funding towards more than just agriculture and home economics (e.g., community and economic development).

The original clientele of the Agricultural Extension Service were primarily rural and farm families and their communities. When the Smith Lever Act was enacted in 1914, over one-half of the population lived in rural areas and over 30% were directly involved in farming (USDA NIFA, 2010). Today, the majority (over 90%) of the population resides in urban or suburban areas, and less than 2% of the total population is directly involved in farming (USDA NIFA, 2010). With less traditional clientele available and a reduction in funding, a shift in programming towards all of society was necessary. Furthermore, agriculture in the United States has become much more efficient and productive requiring fewer farmers to supply the nation with an abundant food supply. In addition to a fewer number of farms, farming operations have become much larger in size. Agriculture is more industrialized and concentrated with much vertical integration (Ilvento, 1997). The original philosophy of Cooperative Extension has not changed much, but the clientele and subject matter have evolved and changed with society.

The land-grant universities were founded with the original mission to include teaching, research, and service (Norland, 1990). The purpose of Cooperative Extension is often lumped into the service portion of the mission along with professional service, outreach, and engagement; however, Extension is unique from these other terms in that Extension is also a part of teaching. Extension is teaching which takes the knowledge of the university to the people living throughout the state. The purpose of Extension is
teaching and extending the research-based information off campus and into the communities which is somewhat different from the service portion of the land-grant mission (Norland, 1990). USDA NIFA (2010) reports the purpose of Extension as solving public needs with college or university resources through nonformal and non-credit programs. Extension serves as an unbiased source of information that is uniquely positioned to solve public needs and create partnerships because it is not perceived as a government agency with political agendas, it is not a private business with financial interests, and it is not perceived as a traditional nonprofit organization with its own particular agenda (Patton, 1986). Patton (1986) also stated that “Extension, more than any other organization in modern and future society, sits at the center between the government sector, the public nonprofit sector, the private nonprofit sector, and the private business sector.”

Cooperative Extension is very unique because of its complex funding and accountability structure, and its arrangement varies somewhat from state to state and/or county to county (Franz & Townson, 2008). The National Institute of Food and Agriculture (NIFA) within the United States Department of Agriculture (USDA) is the governing body responsible for the federal partnership of the Cooperative Extension System (USDA NIFA, 2010). NIFA distributes federal formula funds to the land-grant colleges/universities, and provides program leadership to help identify and address current problems and issues. As a result of this complex public funding structure, Cooperative Extension is required by law to be accountable to the public as a federal, state, and community organization while serving its mission to provide university
outreach through effective educational programming. In addition, Cooperative Extension may receive additional funding and accountability requirements from funders when grants or contracts are awarded from private industry and/or private groups, or user fees are charged to clientele (Stup, 2003). Extension often builds working and/or funding partnerships with both government and non-government organizations (Franz & Townson, 2008). Commodity or industry groups have been traditional partners with Extension and often serve as both funding source and program beneficiary (Aguilar & Thornsbury, 2005).

Ellen Taylor-Powell and Heather H. Boyd (2008) have identified three central pressures for conducting evaluation in Cooperative Extension. The first central pressure is the demand for accountability and documentation of evidence of impact as required by the Government Performance and Results Act (GPRA) of 1993. Also included in this first pressure are additional accountability requirements by state, local, private, and individual funders. The second is internal pressure from Extension itself to become a learning organization with evaluation as a core function for the improvement of programs. Evaluation should be seen as an organizational responsibility, and not a separate or voluntary duty. The third central pressure relates to providing evidence of scholarship as part of the university faculty. Extension personnel desire to produce scholarly work that is valued by peers and external reviewers. In summary, Extension personnel must conduct evaluation to show accountability, to create effective programming (effectiveness/improvement), and to fulfill the mission of the land-grant university (evidence of scholarship).
As a recipient of public funding, the law clearly shows that Extension programming must be accountable to the taxpayers (United States Government, 1993). It is also obvious through a review of the literature in the Journal of Extension that accountability continues to become more and more important as funding and resources become increasingly limited (Kelsey, 2008; O’Neill, 1998; Decker & Yerka, 1990). However, accountability should not be the only primary purpose for evaluation in Extension. Decker and Yerka (1990) documented a statement from a state program leader that evaluations should be 75% useful to the programmer (i.e., improvement/effectiveness) and 25% useful for administrative reporting needs (i.e., accountability). Accountability could be satisfied if a program is judged to be acceptable by funders and/or participants through a simple review of what happened, but effectiveness and evidence of scholarship requires something more substantial.

Kelsey (2008) identified a need for additional educational training on evaluation methods for Extension personnel. Guion, Boyd, and Rennekamp (2007) found that on-the-job experience and independent study were the greatest influence in Extension educators’ knowledge of evaluation. Cooperative Extension is working to build evaluation capacity of faculty and staff and has created an Extension Education Evaluation Topical Interest Group (EEE-TIG) under the auspices of the American Evaluation Association (Guion, Boyd, & Rennekamp, 2007). This is important if Extension personnel are going to conduct higher level evaluations that provide evidence of program impact. Many Extension personnel consider programs successful in lower levels of the program progression (Richardson, 2005). This means that even if a program
has produced tremendous impact, too often the appropriate data is not collected to
document results and provide evidence for accountability and justification for further
funding. Furthermore, the true effectiveness may never be realized and, thus, not used to
improve or continue future related programming.

Background – Program Evaluation

Society faces challenges in which public and nonprofit organizations attempt to
address or resolve through programming. Administrators must make a variety of
decisions regarding the improvement of programs and/or the continuance of programs.
In order to intelligently make decisions, good information on effectiveness is needed for
administrators and policy makers (Fitzpatrick, Sanders, & Worthen, 2003). Program
evaluation is the process used to gather this information. It is more formally defined as
the process of focusing on questions and topics of concern, collecting appropriate
information, and then analyzing and interpreting the information for a specific use and
purpose (Taylor-Powell, Steele, & Douglaah, 1996). Evaluation is defined as the
identification, clarification, and application of defensible criteria to determine an
evaluation object’s value (worth or merit) in relation to those criteria (Fitzpatrick et al.,
2003). In program evaluation, the evaluation object is typically an educational or social
program. The criteria are determined by the evaluators and/or other stakeholders.

The actual practice of program evaluation initially became very important during
the 1960s when the U.S. federal government began allocating substantial funding and
resources for social programs (Fitzpatrick et al., 2003; Scriven, 2003). The federal
government wanted to determine the effectiveness of social initiatives, improve accountability to the American people, and legitimize newly created public programs (Carman, Fredericks, & Introcaso, 2008).

The two major types of program evaluation are formative and summative. Formative evaluation is used for purposes regarding improvement; and summative evaluation is used for purposes regarding adoption, continuation, or expansion (Fitzpatrick et al., 2003). The distinction between formative and summative evaluation is important in determining how the results will be used or what kinds of judgments, decisions, or choices will be made; but it does not determine the nature of the evaluation questions. The nature of the evaluation questions are described by the terms of needs assessment, process or monitoring studies, or outcome studies. Needs assessments are concerned with identifying problems and recommendations for solutions and deal with the situation and priorities. Process studies tend to focus on how the program is being delivered. Outcome studies are concerned with describing, exploring, or determining changes that occur in program recipients, secondary audiences, or communities as a result of the program (Fitzpatrick et al., 2003). As a result of outcome studies, evaluators might use findings to create a simple report of what activities occurred or how many people participated; or they might produce a detailed report with hard evidence showing impact and greater benefits to society.

The primary purpose of program evaluation has evolved to include both accountability and performance. The increased emphasis on accountability resulted from the GPRA of 1993. This act was intended to improve federal program effectiveness and
public accountability by promoting a new focus on results, service quality, and customer satisfaction by systematically holding federal agencies accountable for achieving program results (United States Government, 1993). Program administrators must be accountable to those who are providing the funding and resources (public or private). In today's economic climate, public and private entities are critically evaluating expenditures as they search for more efficient and effective allocation of available resources (Aguilar & Thornsbury, 2005). Accountability becomes more important as financial support becomes less available (Chapman-Novakofski et al., 1997). Richards (2001) stated that government programs must have sufficient public benefits that make them worthy of continued public financial support.

The other primary purpose of program evaluation involves performance or improvement of programs. Not only must programs be held accountable, but programs must also be relevant and accomplish something tangible. Richards (2001) stated that program relevance to constituents and society is now the public norm for determining the value of both public and private organizations. Public programs must provide public benefit or people impacts (financial gains, taxpayer savings, efficiencies gained, environmental enhancements or protection, individual life enhancements, resources preserved, or societal improvements) to show program accomplishments (Bennett, 1996). Program evaluation allows administrators to determine their program’s value, merit, or worth; and to make improvements so that programs are both accountable and accomplishing desirable end results. A recent study investigating the factors that affect Extension practitioners and their behavior regarding program evaluation found that many
respondents were more satisfied when their evaluation results were used for more than just for the sake of accountability (Morford, Kozak, Suvedi, & Innes, 2006).

Cooperative Extension provides nonformal education; therefore, it is important to note that conducting program evaluation in nonformal educational setting is somewhat different from that required to take place in more formal educational settings. Nonformal programs are typically offered by groups other than traditional educational institutions; and the structure, support, enthusiasm, and capacity for conducting program evaluation tends to be less (Norland, 2005). Practitioners working in nonformal programs are often trained in fields outside of education and are not educators by formal training (Wiltz, 2005). The program’s decision makers are typically members of a board of directors or advisory boards with various backgrounds. They may be philosophically removed from the educational component and more focused on the bottom line such as finances or efficiencies (Norland, 2005). Nonprofit organizations working outside of public oversight might be unfamiliar with evaluation and even fearful of it (Wiltz, 2005).

Furthermore, nonprofit organizations providing nonformal education often work on a fixed budget dependent on outside sources of funding; thus, program evaluation could be viewed as an additional and unnecessary expense.

The mission and funding sources of Cooperative Extension are often somewhat similar to that of nonprofit organizations. Extension could be described as a quasi-nonprofit as portions of this very complex system are voluntary and separate from government. The activities of Cooperative Extension are conducted for not-for-profit purposes; although, educational programming may result in increased profits for
participants. Evaluation of a private organization can be as simple as examining the bottom line of a profit and loss statement; however, evaluation of success in a mission-driven nonprofit organization is much more difficult (Sawhill & Williamson, 2001).

Nonprofits and religious organizations provide numerous public services and typically receive both public and private funds to provide these services (Carmen, Fredericks, & Introcaso, 2008). Society invests in nonprofit organizations because it believes in the organization’s mission. However, the mission of the organization is simply words or intensions. It is the actual programs, activities, or actions of the organization that make a difference for society. The activities of the organization are carried out in service of the mission; therefore, there should be a clear and strong link between evaluation and mission achievement (Hoole & Patterson, 2008). Funders typically demand accountability, which means that nonprofits must conduct program evaluation and submit data or findings. Hoole and Patterson (2008) pointed out that for program evaluations to be meaningful and useful; funders must actually use the evaluation data to make decisions. Hoole and Patterson (2008) used previous publications to support their argument that demonstrating impact is difficult leading nonprofits to gather useless and sometimes misleading data. The point was that evaluation is simply an exercise when the sole purpose was to meet the accountability requirements of stakeholders and it was not tied to making an impact that supported the organization’s mission and the people/issues in which the organization serves (Hoole & Patterson, 2008).
Background – Impact

The results of an evaluation should not simply be a review or account of what happened as the funding source often wants to see measures of program effectiveness in addition to descriptive reports of activities (O’Neill, 1998). In Extension, the evaluation should not only determine what resulted from the program, but determine the impact of the program. Impact can be positive or negative. In program evaluation, the term is typically used for describing something positive. Impact represents a higher level of results or accomplishments. Keith Diem (2003) defined impact in Extension, “as the positive difference we make in people’s lives as a result of the programs we conduct”. The results of a program must ultimately change people’s attitudes or behavior, or benefit society in other ways (Diem, 1997).

Diem (2003) identifies five reasons for documenting impact. The first is to justify the investment of time and effort, as well as the dedication of public and private funds. The second is to earn and build professional, organizational, and political credibility and support. The third is to satisfy the requirements of political bodies and funding agencies. The fourth is to yield tangible results that serve as a basis for scholarly publications, as well as awards and recognition. The fifth requires determining to what degree participants achieve intended results. Richardson (2005) believes that impact should include substantive social, economical, or environmental benefits to both clientele and the general public.

Program evaluation in Extension must document impact to satisfy the evaluation objectives of accountability and performance. After citing supporting articles published
in the *Journal of Extension* as evidence, Richard Stup (2003) stated that Extension must increase evaluation efforts in order to generate systematic and convincing evidence of programming value. His intention was to not only show that better documentation of value will help with the continuation of funding, but that additional funding can also be acquired from external sources by providing programming with documented value.

**Background – Evaluation Tools for Documenting Evidence of Impact**

There are countless evaluation models and tools utilized for program evaluation in Cooperative Extension; however, Claude Bennett (1975) developed a hierarchy for documenting levels of evidence of program impacts in Cooperative Extension. His seven-link “chain of events” included: inputs; activities; people involvement; reactions; knowledge, attitudes, skills, and aspirations (KASA) change; practice change; and end results. Inputs include the resources invested by Extension such as time, money, and staff. The inputs produce activities (e.g., meetings, sessions, workshops, or events) involving people who have reactions, both good and bad. In a productive program, the participants may change their KASA as a result of the programming. Practice change occurs if the changes in knowledge, attitudes, skills, and aspirations are actually applied. End results are the broader outcomes, effects, and benefits to the community or society such as increased profits, better health, or cleaner environment. End results help accomplish the overall mission of CES such as positively affecting social, economical, civic, and environmental conditions for society.
In Bennett’s Hierarchy (also referred to as Bennett’s Hierarchy of Evidence), true impact increases as one goes up the hierarchy as the lower few levels are important precursors, but are not necessarily evidence of impact (Diem, 2003). Evaluation at any of the seven levels may help in decision making, but actual impact typically occurs above the level of simply involving people. Evidence of more significant impact would be measured at the levels of practice change and end results (Diem, 2003). Also, evaluation is more difficult and time consuming as the levels increase (Bennett, 1975; Diem, 2003).

The modern logic model, in which Bennett’s Hierarchy was a forerunner, has achieved widespread adoption within Cooperative Extension during the past 15 years (Taylor-Powell & Boyd, 2008). The term “logic model” was first mentioned in the 1960s. The publications produced by the United Way of America in 1996 and the W. K. Kellogg Foundation in 2001 significantly helped to increase its popularity and use worldwide (Knowlton & Phillips, 2009). It has become very popular today across a wide variety of organizations because of its multifaceted value in helping people plan, implement, communicate, and evaluate programs. The logic model is a systematic and visual way to present and share your understanding of the relationship among resources you have to operate your program, the activities you plan, and the changes or results you hope to achieve (W. G. Kellogg Foundation, 2004). The logic model terminology, components, and structure may vary somewhat across organizations and individuals; therefore, the University of Wisconsin-Extension (UWEX) logic model format will be used for purposes of this study.
The University of Wisconsin-Extension logic model format has evolved since 1995 when it was originally created in response to the GPRA of 1993 (Taylor-Powell & Henert, 2008). It was specifically developed for use by Cooperative Extension based on Bennett’s Hierarchy and the USAID Log Frame (Taylor-Powell & Henert, 2008). This logic model format contains the following components: situation/priorities, inputs, outputs-activities, outputs-participation, short-term outcomes, medium-term outcomes, long-term outcomes, assumptions, and external factors. When documenting impact, outcomes are much more important and provide much stronger evidence than outputs. Short-term outcomes are considered to be immediate such as changes in awareness, knowledge, skills, attitudes, opinions, and intent (Taylor-Powell & Henert, 2008). Medium-term outcomes are considered to be intermediate such as changes in behaviors, decision-making, and actions (Taylor-Powell & Henert, 2008). Long-term outcomes are the final outcomes (often referred to as impact) such as changes in social, economic, civic, and environmental conditions (Taylor-Powell & Henert, 2008).

It is important to note that much of Bennett’s Hierarchy is embedded within the logic model. For example, the University of Wisconsin-Extension logic model format directly shows inputs, activities, and participation. KASA change would be considered as short-term outcomes, behavior change as medium-term outcomes, and end results as long-term outcomes. Reaction is the only aspect of Bennett’s Hierarchy that is not directly represented within this logic model format; however, satisfaction is typically included as a part of participation in the logic model and can be considered as a form of reaction. Just as evidence of impact increases as one goes up Bennett’s Hierarchy,
Both Bennett’s Hierarchy and the logic model are very helpful tools for collecting and documenting evidence of impact. In addition, they both work extremely well for labeling and describing a specific level of evidence of program impact. The recent widespread use and adoption of the logic model in Cooperative Extension (in place of Bennett’s Hierarchy) is likely a result of the logic model’s diverse usefulness for program planning and grant writing purposes in addition to evaluation reporting (Taylor-Powell & Boyd, 2008). Bennett’s Hierarchy is important to this study because one can argue that it is the simplest, easiest to understand, and contains the most applicable description of different levels of evidence of impact. The University of Wisconsin-Extension logic model format is important to this study because one can argue that it is the more contemporary and predominant tool currently used when documenting and describing impact resulting from Extension programming.

Background – Reach, Richness, & Geographic Scope

Extension programming can include a wide array of activities ranging from a simple service to very in-depth educational training. It can be said that typically Extension either takes a very large number of people a small step and delivers very general information (e.g., fact sheets), or Extension takes a very small number of people a large step and delivers very in-depth information/training (e.g., short course). To describe this assertion, David King and Michael Boehlje (2000) referred to Evans and
Wurster’s concepts of “reach and richness” and applied them to the delivery of information by Cooperative Extension. King and Boehlje (2000) defined reach as the number of individuals in which the interaction takes place (size of audience), and richness as the detail and depth of this interaction. Additional definitions provided by Evans and Wurster (2000) refer to reach as the number of people who participate in the sharing of the information, whereas richness refers to the quality of information as defined by the user.

David King and Michael Boehlje (2000) used the concepts of “reach and richness” to make the point that for Extension to remain relevant, personnel should become more focused on increasing reach (without losing richness) or private industry will attempt to fill the void. If unbiased Extension programming is not making a significant impact, then private industry (potentially biased) will become the predominant source of information and programming for current and future Extension clientele. “If you can’t reach enough people with your message, you can’t capture enough value (or impact) to continue to be viable; however, if you can compress the process and reach more people with a greater degree of richness, you have the opportunity to capture significantly more value (or impact) (King & Boehlje, 2000).”

The concepts of “reach and richness” are described as the economics of information. Evans and Wurster (2000) have stated, “to the extent that information is embedded in physical modes of delivery, a basic law governs its economics: there is a universal trade-off between richness and reach.” Historically, this relationship has been considered as a continuum; thus, you must sacrifice one to achieve the other (King &
Boehlje, 2000). Delivering very rich information has required proximity and dedicated channels of delivery, and increasing the reach of this very rich information has been impacted by costs and physical constraints (Evans & Wurster, 2000). The concepts of “reach and richness” are important to this study because even though a high level of impact might be documented by behavior change and application of knowledge, the overall benefit of that impact depends on the significance of the number of people reached and the degree of richness.

Another variable which may also affect the overall benefit of impact is the geographic scope which is defined as the geographic region/area or location of the clientele/audience targeted for programming. The effect of geographic scope on impact is somewhat similar to that of Evans and Wurster’s concept of reach; however, the definitions of geographic scope and reach are very different. Documenting evidence of high impact from a single local program might be of less value or have less overall benefit to society than high impact resulting from a single multistate or national program. However, a program with a smaller geographic scope can have very high richness and make a very significant impact on the local community. This program can then be expanded or replicated to benefit a much larger geographic area. Although the effect is somewhat related to reach, identifying any potential relationship between geographic scope and impact is important to this exploratory study because Extension programming has been traditionally developed and targeted towards addressing local needs. However, Extension programming has become more diversified in regards to the geographic scope
of programming. Program structures are now planned and developed for county (local), multicounty regional, statewide, multistate regional, or national clientele.

**Statement of Problem**

Too often, Extension programs which utilize program evaluation in the form of outcome studies do not document impact by collecting real evidence of behavior change and/or greater end results that benefit society through social, economic, civic, or environmental improvement. Traditionally, the criticism of evaluation in Cooperative Extension has been that evaluators have simply reported on the number and demographics of the people who participated (Gentry-Van Laanen & Nies, 1995). Most program evaluation simply stops at the reaction level (reports on inputs, activities, people involvement, and reactions) or learning (KASA change) without measuring higher level changes (Stup, 2003).

Previous research published by Michael Duttweiler (2008) suggested that 35% of all articles pertaining to program evaluation and published in the *Journal of Extension* from 1998 through 2007 provided some evidence of effectiveness according to Jacob’s Five-Tiered Approach. However, it is still not known what levels of evidence of program impacts were documented by outcome evaluation studies according to Bennett’s Hierarchy or the logic model. In addition, there was no information regarding the trend over time as this study examined a single decade (1998-2007). Evidence of effectiveness in Jacob’s Five-Tiered Approach was defined as using experimental or quasi-experimental methodology, and measuring impact using random assignment or
comparison groups. Therefore, Duttweiler’s findings regarding effectiveness showed that advantageous scientific methods were used to document impact for 35% of the applicable articles, but did not indicate what level of evidence of impact was documented (e.g., behavior change or end results [social, economic, civic, or environmental change]). Duttweiler’s study did not investigate or report on any potential trends over time. Furthermore, it is unknown if there is any relationship between the documentation of evidence of impact with reach and richness or geographic scope.

Extension must conduct higher level program evaluations and document outcomes that provide accountability to funders and provide information for program improvement and performance resulting in impact. By addressing the accountability issue, Extension justifies its existence, relevance, and ultimately its value for continued and/or additional funding. By improving programs, even those that already have high impact, Extension better uses research-based knowledge to change people’s behaviors and to produce end results which benefit society through social, economic, civic, or environmental improvement. Furthermore, by showing evidence of impact from individual programs, the overall impact of the Cooperative Extension System and its mission increases and provides more evidence of accountability and effectiveness. Extension programs need to document evidence of higher level impact demonstrated by behavioral changes that ultimately lead to end results which are beneficial to society.
Purpose of the Study

The purpose of this study is to explore and describe the level of evidence of impact collected through program evaluation (outcome studies) by the Cooperative Extension System as published in the *Journal of Extension* (JOE); and to examine any potential relationships between the documentation of impact with reach, richness, and geographic scope. Impact is operationally defined as program effectiveness determined by the highest level of evidence collected based on the components of both Bennett’s Hierarchy and the University of Wisconsin-Extension logic model format. This study will consist of a thorough review of the articles utilizing program evaluation that were published in the *Journal of Extension* under the headings of ‘Feature Articles’ and ‘Research in Brief’ from 1965-2009, consisting of five separate five-year periods (1965-69, 1975-79, 1985-89, 1995-99, and 2005-09).

Objectives of the Study

The objectives of this study were as follows:

1. To determine the number of program evaluations with outcome data published in the *Journal of Extension* during the years investigated and the program areas represented in these evaluation studies.

2. To review a subset of articles published in the *Journal of Extension* from 1965-2009 in order to determine the number and percentage of evaluation studies that reached each level of evidence of impact according to both Bennett’s Hierarchy and the logic model.
3. To investigate and identify any trends over time.
4. To identify and quantify the primary evaluation and data collection methods used to document evidence of impact.
5. To determine if the program being evaluated was designed for reach (number of people interacted with) and/or richness (detail and depth of interaction).
6. To determine the geographic scope (county [local], multicounty regional, statewide, multistate regional, or national) of the programs evaluated.
7. To explore the relationship between reach/richness, geographic scope, and the level of impact according to both Bennett’s Hierarchy and the logic model.

Conceptual Framework

The conceptual framework in Figure 1 shows the progression towards the ultimate goal of impact by the various levels of both Bennett’s Hierarchy and the components of the University of Wisconsin-Extension logic model format. Documenting program outcomes and providing greater evidence of impact is the strongest form of program evaluation, and higher levels of evidence demonstrate greater impact. The independent variables would be Level-Bennett’s Hierarchy and Level-UWEX Logic Model. Each of the individual levels or components is a different level of the independent variable. Impact is the dependent variable.

The moderating variables which potentially affect the overall quality or benefit of that impact (dependent variable) are reach, richness, and geographic scope. If reach is low and very few people are impacted, the impact might be of less overall value or
benefit to society. Vice versa, if the information was not very rich or in-depth, it might be of less overall value or benefit to society. Geographic scope works very similar to reach. If an individual program is targeted towards a single county, it might be of less overall benefit to society than a program targeted towards an entire state, multiple states, or the nation. It has been previously understood that reach and richness tend to work against each other like a continuum unless physical modes of delivery are removed. It is currently unknown as to how geographic scope relates to impact.

**Figure 1: Conceptual Framework**
**Definition of Terms**

The following terms are defined both constitutively and operationally as they were used in the context of this study.

- **Evaluation**
  - The identification, clarification, and application of defensible criteria to determine an evaluation object’s value (worth or merit) in relation to those criteria (Fitzpatrick et al., 2003).
  - Operationally defined as any work that involved “the systematic acquisition and assessment of information to provide useful feedback about some object” (Trochim, 2006).

- **Program evaluation**
  - The thoughtful process of focusing on questions and topics of concern, collecting appropriate information, and then analyzing and interpreting the information for a specific use and purpose (Taylor-Powell, Steele, & Douglah, 1996).
  - Operationally defined as evaluative outcome studies published as articles in the *Journal of Extension* which expressly referenced the acquisition and use of programmatic information.

- **Program**
  - A system of projects or services intended to meet a public need.
    
    Extension programs may include teaching, published curriculum,
volunteer training, or applied research and may or may not involve the public directly while they are being delivered (Diem, 2003).

- Operationally defined as Extension educational programming including individual programs, program areas, and program systems.
  - **Individual programs** are considered to be pieces of programming that Extension personnel teach or facilitate (e.g., workshops, training, short courses, camps).
  - **Program areas** are considered to be the statewide units of Agriculture and Natural Resources; 4-H Youth Development; Family and Consumer Sciences; and Community and Economic Development.
  - **Program system** is considered to be the State or U.S. Cooperative Extension System.

- **Educational Tool**
  - Anything used as a means of accomplishing an educational task or purpose.
  - Operationally defined as Extension educational products such as: newsletters, bulletins, web sites, television or radio programs, etc.

- **Impact**
  - The positive difference Extension makes in people’s lives as a result of Extension programs. The results of a program must ultimately change
people’s attitudes or behavior, or benefit society in other ways (Diem, 1997; Diem, 2003).

o Operationally defined as program effectiveness determined by the highest level of evidence collected based on the components of both Bennett’s Hierarchy (e.g., inputs, activities, people involvement, reactions, KASA change, practice change, and end results) and the University of Wisconsin-Extension logic model format (e.g., situation/priorities, inputs, outputs-activities, outputs-participation, short-term outcomes, medium-term outcomes, and long-term outcomes).

• Reach

  o How many individuals you can touch and influence (King & Boehlje, 2000). Also, the number of people who participate in the sharing of that information (Evans & Wurster, 2000).

  o Operationally defined as either high reach or low reach.

    ▪ High reach programs are those that (1) are created and targeted to reach a larger audience, (2) may not require any formal registration or attendance, and/or (3) may not contain any evidence or record of participation by any single individual.

    ▪ Low reach programs are the exact opposite with criteria consisting of (1) programs created and targeted to reach a smaller and more well-defined audience, (2) programs that often require
formal registration and/or attendance, and/or (3) programs with evidence of participation and attendance.

- **Richness**
  
  - The detail and depth of your interaction (King & Boehlke, 2000). Also, the quality of information as defined by the user: accuracy, bandwidth, currency, customization, interactivity, relevance, security, and currency (Evans & Wurster, 2000).
  
  - Operationally defined as either high richness or low richness.
    
    - **High richness** programs are those that provide very in-depth information and typically require participants to attend or participate multiple times allowing them the opportunity to practice and apply their new knowledge throughout the program’s duration. Criteria for high richness consists of (1) high bandwidth in which very in-depth information is moved from sender to receiver, (2) high customization in which the program is specifically tailored to a specific audience, (3) high interactivity in which there is a great deal of exchange between program providers (e.g., instructors, administrators) and participants, (4) high reliability in which participants tend to trust program providers, (5) high security in which sensitive information could be shared, and/or (6) high currency in which information exchange (feedback) is immediate.
- **Low richness** programs are those that are typically one shot programs with (1) low bandwidth in which very general information is provided, (2) low customization in which the program is not specifically targeted to a specific audience, (3) low interactivity in which there is little to no exchange between program providers (e.g., instructors, administrators) and participants, (4) low reliability with less trust as participants and program providers are likely strangers, (5) low security in which sensitive information is unlikely to be shared, and/or (6) low currency in which information exchange (feedback) is slower.

- **Geographic Scope**
  - The geographic region/area or location of the clientele/audience targeted for the programming.
  - Operationally defined as county (local), multicounty regional, statewide, multistate regional, or national based on the program represented in the article/study.

**Limitations of the Study**

There are some limitations that were recognized for this study. First, many program evaluations conducted for Extension programs are likely published and printed in various other journals and publications. The *Journal of Extension* was selected for the purposes of this study because it is reported to be the official refereed journal of the U.S.
Cooperative Extension System (Extension Journal, Inc., 2010). Another aspect related to this limitation involves examining only those outcome evaluation results that were actually published. It is possible that many programs were formally evaluated with findings that were utilized for accountability and/or improvement, but the results were never formally published. These unreported program evaluations may have included the documentation of evidence of high level impact.

A second limitation results from the fact that only articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’ where reviewed for this study. Additional articles reporting on outcome studies and the documentation of evidence of impact might be included under other headings, specifically ‘Ideas at Work’ and ‘Tools of the Trade’; but these articles are typically shorter and less comprehensive, often missing pertinent information required for the purposes of this study. Another aspect related to this limitation is that the researcher is also limited to the actual information contained within the individual articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’. Some authors may tend to give a much less comprehensive description of their program and their evaluation efforts.

A third limitation involves the Journal of Extension itself. The journal has evolved and changed policies and procedures throughout its existence. There were likely different requirements for publishing articles in the 1960s than in the 2000s. Today, the types of articles published under the categories of ‘Feature Articles’ and ‘Research in Brief’ are somewhat similar; however, in the 1960s the articles contained in ‘Research in Brief’ tended to be abstracted by the JOE editor from other journals or from an
unpublished thesis or dissertation. A similar limitation involves the PDF of the article as presented on the JOE web site. Sometime the scanning quality was poor and a portion of the article appears very black or distorted. This required more careful attention and educated guesses in reading the distorted words.

Finally, an additional limitation involves the clientele who publish information in the Journal of Extension. Agent/Educators and other Extension personnel may or may not have a tenure-track appointment with publication requirements. Those without publication requirements might be less likely to publish evaluation results. Those specialists located on campus with tenure-track appointments may be more apt to publish in other research journals more specific to their field of expertise. State Extension Specialists are often responsible for major statewide programs which could be more likely to be formally evaluated with published results than smaller, localized programs.

Basic Assumptions

For purposes of this study, the researcher made the assumption that abstracts discarded in the screening process accurately reflected the content of the article and those respective articles discarded did not report evaluative results from an outcome study. However, professional judgment was used to scan further into an article if the abstract was extremely short or didn’t exist. In addition, it was assumed that the highest level of evidence of impact documented for each program evaluation was reported in the individual published article. For example, it is assumed that the authors would report on evidence of end results/long-term outcomes if they had actually collected this higher level
evidence, and not simply reported on lower levels such as KASA change/short-term outcomes.

**Significance of Study**

This study is exploring the areas of accountability and performance which are important to Extension personnel and administrators. It is generally believed that providing documented evidence of impact benefitting society will influence legislators to provide continued funding or additional support. If this is the case, Extension personnel need to know what percentage of programs are currently measuring and documenting impact. They also need to know what percentage are currently only reporting lower (or insignificant) levels of evidence of impact such as the number of people attending and/or the reactions of those attending.

It is currently unknown as to how many program evaluations reported in the *Journal of Extension* are collecting evidence and making significant impact that benefits society. Furthermore, by identifying and quantifying the various methods used to document higher levels of evidence of impact (i.e., practice change, end results, medium-term outcomes, and long-term outcomes) other program leaders may benefit by utilizing similar methods. Many have uttered the phrase that “too often program evaluation simply involves passing out a ‘smile sheet’ at the end of a program to find out how much respondents liked or disliked the program”. Is this really true for program evaluations that are published in the *Journal of Extension*? The results of this study will answer this and other important questions.
Furthermore, understanding the current relationship of impact with reach, richness, and geographic scope might help Extension personnel with future program development. Should more resources of time, labor, and money be spent on more in-depth programming for fewer people? Should resources be targeted to as many people as possible using more mass media methods of delivery?

Another aspect similar to reach that might be beneficial to Extension personnel when documenting impact involves the geographic scope. Extension personnel will better understand where opportunities may exist for creating new programming by determining the relationship between geographic scope (county [local], multicounty regional, statewide, multistate regional, or national) and the level of evidence of impact documented. Personnel may also identify what types of programs produce the greatest impact. For example, if multistate programs appear to demonstrate higher levels of impact more frequently, then more future resources should be dedicated to multistate efforts (or, more program evaluation efforts should be dedicated to county programs).

Understanding the present situation regarding reach, richness, geographic scope, and the documentation of evidence of impact may also be helpful in creating new sources of funding in terms of grants, contracts, fiscal gifts, or user fees.

To summarize, the findings of this study will help Extension personnel better understand the current usage of program evaluation and the documentation of evidence of impact. Personnel will also be able to use the results to identify shortcomings and improve their program evaluation efforts. It is possible that certain methods or procedures will be identified as more relevant, more successful, or more frequently used.
to document certain levels of impact. Those methods or procedures found to be the most successful could be used in much greater frequency for future program evaluations to strengthen the evidence collected and better demonstrate impact.
CHAPTER 2
REVIEW OF LITERATURE

The purpose of this literature review is to provide a description and historical context of using program evaluation for the documentation of evidence of impact within the Cooperative Extension System. Countless evaluation approaches and models are utilized to evaluate Extension programs; however, Kirkpatrick’s model (1959), Bennett’s Hierarchy (1975), and the logic model (1995) are visual displays that help to document and describe various levels of evidence of impact. In 1975, Bennett’s Hierarchy was introduced and published for the first time in the *Journal of Extension* specifically for use by Cooperative Extension personnel. Bennett’s Hierarchy was created by incorporating the elements of Donald Kirkpatrick’s four-levels (introduced and published during 1959) for evaluating training programs (Rockwell & Bennett, 2004). Another modern or more recent tool that has progressed from both Kirkpatrick’s model and Bennett’s Hierarchy is the logic model (Hoffman & Grabowski, 2004).

Today, Extension personnel are potentially more familiar and more likely to refer to the logic model when publishing the results of a program evaluation. Although the logic model is not exactly considered an evaluation model, it is a tool that allows program administrators to visualize what should be evaluated based on the level of evidence of impact they wish to measure. For example, KASA change might be measured as a short-
term outcome, behavior change as a medium-term outcome, and end results (economic, social, civic, and/or environmental change) as a long-term outcome (Taylor-Powell & Henert, 2008). Because of the current popularity and widespread use of logic modeling, it is important to connect it with the terminology used to describe different levels of impact in Bennett’s Hierarchy.

This chapter begins by introducing Ewert’s model which demonstrates the diversity and complexity of conducting program evaluation in the Cooperative Extension System. The delivery method used and information provided from Extension programming varies significantly; thus, affecting the value of the impact of programming. This chapter will then move into a description of the individual evaluation tools or models utilized for the actual documentation of impact. The literature review provides an examination of the progression of program evaluation and the documentation of impact in Cooperative Extension, specifically focusing on visual displays (or models) to document impact such as Kirkpatrick’s model (1959), Bennett’s Hierarchy (1975), and the logic model (1995). In addition, an examination of the literature regarding reach and richness and how it relates to Extension programming will be included as well as a look at the geographic scope of programming.

Conducting Program Evaluation in Extension

Conducting program evaluation in the Cooperative Extension System can be a challenging endeavor because of the diversity and complexity of the organization. There are usually multiple different funding sources and tremendous diversity in the delivery of
programming and the type of information. Evaluations may range from an individual local program with very few participants to a major statewide evaluation of an entire state’s Extension system. In evaluating Extension programming, it is important to consider the domain of the programming because Extension personnel provide such a variety of services and diverse education. Franz and Townson (2008) described the model created by Merrill Ewert, a past director of Cornell Cooperative Extension.

Ewert’s model included four domains for Extension work which focused on process as the delivery method and content as the information delivered (Franz & Townson, 2008). The four domains shown in Figure 2 include service, facilitation, content transmission, and transformative learning. Programs falling into each of these domains should be evaluated to measure impact, but typically overall impact will be greater in the domains with higher process and/or content. However, a program falling into the service domain (lower process and content) can still be evaluated and found to produce high impact if addressing a critical need.

The primary purpose of Extension is not to simply provide service, but to teach and extend the research-based information from the land-grant universities into the communities (Norland, 1990). More or stronger evidence will typically be needed to show evidence of higher-level impact for low process and content (service domain). Various models or tools are utilized to document impact for programming falling into each of these domains. Kirkpatrick’s Model was one of the first tools used for describing levels of evidence of impact for training programs.
Kirkpatrick’s Four-Level Model of Evaluation

Much of the educational programming in Extension could be considered somewhat similar in type and kind to the programming that businesses use for training employees as Extension is essentially training individuals in life. Typically, business managers are primarily interested in how their training has changed their employees’ behaviors or performance and not simply their employees’ reaction to that training (Stup, 2003). Business managers aren’t necessarily concerned whether participants liked, disliked, or enjoyed a training program. They might be somewhat interested in knowing that learning has actually taken place, but they would be more so interested with changes in behavior and changes in performance which ultimately results in increased production/profits (Stup, 2003).

Kirkpatrick’s model has served as the standard framework for evaluating training programs (Rajeev, Madan, & Jayarajan, 2009; Bates, 2004; Holton, 1996). It is well

Figure 2: Model for Extension Education Approaches created by Dr. Merrill Ewert, past director of Cornell Cooperative Extension (Franz, 2008)
adapted for providing the kind of information desired by business managers (Stup, 2003). The model, originally developed and published by Donald Kirkpatrick in 1959, has also been used in Cooperative Extension as both a standalone evaluation model and as a foundation for newer models (Bennett, 1975; Hoffman & Grabowski, 2004). The Kirkpatrick Model is important to the present study because it was the initial model developed to examine more tangible measures of impact instead of simply measuring the reactions or feelings of the participants, and it has served as the basis and forerunner for Bennett’s Hierarchy and the modern-day logic models.

The Kirkpatrick Model as shown in Figure 3 consists of four levels (or steps) including: (1) reaction, (2) learning, (3) behavior (or transfer), and (4) results (Kirkpatrick, 1959, 1996, 2010). Reaction is a measure of how participants feel about the program. Learning is a measure of the knowledge acquired, skills improved, or attitudes changed as a result of the program. Behavior is a measure of the extent to which participants changed their on-the-job behavior as a result of the program. Behavior may also be referred to as transfer or application (Kirkpatrick, 1996; Winfrey, 1999). Results, the fourth and final level, is considered to be the final results or performance change which occurred because of the program such as increased quality, productivity, or profits. Kirkpatrick (1959, 1996) reports that evaluation becomes more difficult, complicated, and expensive as the levels increase; however, the evaluation becomes much more meaningful.
The Kirkpatrick Model has been criticized for the lack of research to further develop a theory of evaluation. Holton (1996) believed that it was simply a classification scheme or taxonomy, and that higher level research was needed to create a theory supported by scientific evidence. Reid Bates (2004) identified three major limitations of the Kirkpatrick Model. The limitations were the incompleteness of the model, the assumption of causality, and the assumption of increasing importance of information as the levels are ascended (Bates, 2004). Bates (2004) also reported that the model was unable to address both the summative purposes (was training effective?) and the formative purposes (how can training be modified in ways that increase its potential for effectiveness?) of program evaluation.

Donald Kirkpatrick is not concerned as to whether his four levels are considered a model or taxonomy, but is satisfied that they remain useful to professionals for evaluating their training programs (Kirkpatrick, 1996). As described later in this chapter, the logic model overcomes some of this previous criticism by acknowledging the role of external factors, which include new technologies and social pressures that can slow or accelerate
practice adoption (Hoffman & Grabowski, 2004). Jim Kirkpatrick, which is Donald Kirkpatrick’s son, has published an article which provides evidence demonstrating that the four levels remain very relevant today (Kirkpatrick, 2007). The original work gave three reasons to evaluate. Donald and Jim Kirkpatrick have created three additional reasons to evaluate programs of which two of the reasons addressed the formative and summative evaluation concerns (Kirkpatrick, 2007). The six reasons to evaluate are as follows (Kirkpatrick, 2007):

1. Determine whether a program should be continued. (Summative)
2. Improve a program. (Formative)
3. Ensure learning compliance.
4. Maximize the value of training.
5. Align training with strategy.
6. Demonstrate the value of training.

Bennett’s Hierarchy

Claude Bennett originally developed his hierarchy or seven-link “chain of events” for the evaluation of Cooperative Extension programs in 1975. The levels of the hierarchy as shown in Figure 4 are as follows: inputs; activities; people involvement; reactions; knowledge, attitudes, skills, and aspirations (KASA) change; practice change; and end results. Figure 5 provides an alternative illustration of the various levels.

Bennett’s Hierarchy has many common characteristics with Kirkpatrick’s four levels (Rockwell & Bennett, 2004). However, the hierarchy is much more detailed at the
lower levels as it was designed for Extension personnel who are concerned with all facets of programming instead of business managers who are primarily concerned with the highest levels (i.e., behavior change and performance change) (Stup, 2003). This is an important point for Extension evaluators to realize because evaluations ending at Bennett’s level of people involvement do not even fully address Kirkpatrick’s first level (reactions). Just like the Kirkpatrick Model, evaluation is more difficult and time consuming as the hierarchy increases (Bennett, 1975; Diem, 2003). Traditionally, the criticism of evaluation in Cooperative Extension has been that evaluators have simply reported on the number and demographics of the people who participated (Gentry-Van Laanen & Nies, 1995). Most evaluation simply stops at reaction (i.e., inputs, activities, people involvement, reactions) and learning (i.e., KASA change) without measuring higher level changes (Stup, 2003).

Figure 4: A Hierarchy of Evidence for Program Evaluation (recreated from Bennett, 1975)
In more recent years, Kay Rockwell and Claude Bennett (2004) have worked together to further modify and expand the original hierarchy created in 1975. They have added program development and slightly modified the evaluation portion to create a system called Targeting Outcomes of Programs (TOP). The primary purpose of TOP is

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>End result(changes in condition: changes in the human, economic, civic, biological condition) What is the long term impact of your program? (e.g., increased survivorship of species)</td>
</tr>
<tr>
<td>6</td>
<td>Action: changes in behavior, practice, decisions, policies, social action (e.g., improved legislation around hunting of species)</td>
</tr>
<tr>
<td>5</td>
<td>KASA changes: changes in Knowledge, Attitudes, Skills, and Aspirations (e.g., increased species-friendly land management skills)</td>
</tr>
<tr>
<td>4</td>
<td>Reactions: How did stakeholders react to the program? Were they satisfied? Were their expectations met? Was the program appealing? Do they perceive any immediate benefits?</td>
</tr>
<tr>
<td>3</td>
<td>Involvement: How many stakeholders participated? Who participated?</td>
</tr>
<tr>
<td>2</td>
<td>Activities: What activities were involved (content or subject, method, and techniques)? How do we achieve our objectives? What do we do?</td>
</tr>
<tr>
<td>1</td>
<td>Inputs: What resources were expended on the product (time, money, staff)? How do we achieve our objectives?</td>
</tr>
</tbody>
</table>

Figure 5: Bennett’s Hierarchy- Contemporary Example (Leech, Sutherland & Wainwright, 2004)
to reduce social, economic, and environmental problems and document the progress (Rockwell & Bennett, 2004). The TOP system was designed and targeted for use by any organization interested in addressing problems in society. It integrates program development with program performance and focuses on outcomes in planning, implementing, and evaluating programs. This study will focus on the original Bennett’s Hierarchy, which is essentially the program performance or evaluation portion of TOP. In addition to Bennett’s Hierarchy, this study will also focus on the logic model when describing levels of evidence of impact.

Logic Model

The logic model continues the progression, as the roots of the model are found in both Kirkpatrick’s model and Bennett’s Hierarchy (Hoffman & Grabowski, 2004; Leech, Sutherland, & Wainwright, 2004). There are numerous definitions and descriptions for the logic model. There are also various other terms used to describe it such as: idea map, rich picture, action map, mental model, program action, program logic, conceptual map, outcome map, and model of change (Knowlton & Phillips, 2009; Taylor-Powell & Henert, 2008). The United Way of America (1996) places a great deal of emphasis on outcome measurement and refers to their format as a program outcome model. The Kellogg Foundation (2004) defined the logic model as “a systematic and visual way to present and share your understanding of the relationships among the resources you have to operate your program, the activities you plan, and the changes or results you hope to achieve.” Ellen Taylor-Powell and Ellen Henert (2008) defined the logic model as “a
depiction of a program showing what the program will do and what it is to accomplish.” Lisa Knowlton and Cynthia Phillips (2009) provided several different definitions, but the primary one was that the logic model is “a graphic way to organize information and display thinking.”

The logic model displays the resources and actions that one believes is needed to reach their intended results (W. G. Kellogg Foundation, 2004). It is a series of if-then relationships that helps lead one from resources to impact (Taylor-Powell & Henert, 2008). In its simplest form, the logic model consists of inputs, outputs, and outcomes (Taylor-Powell & Henert, 2008). More detailed or complex logic models typically contain resources/inputs, activities, outputs, outcomes, and impact (W. G. Kellogg Foundation, 2004). For example, the United Way of America (1996) uses a program logic model format that contains inputs, activities, outputs, and outcomes. The University of Wisconsin-Extension logic model format as displayed in Figure 6 contains the following components: situation/priorities, inputs, outputs-activities, outputs-participation, short-term outcomes, medium-term outcomes, long-term outcomes, assumptions, and external factors.
The University of Wisconsin-Extension logic model format will be utilized for purposes of this study; thus, the various components will be explained in greater detail. The situation is the originating program or issue set within a complex of sociopolitical, environmental, and economic circumstances (Taylor-Powell & Henert, 2008). The inputs are the resources and contributions that go into or are invested into the program (Taylor-Powell & Henert, 2008). Outputs are what the program does and who it reaches such as activities, services, events, products, and people reached (Taylor-Powell & Henert, 2008). Outcomes are the results such as the value or changes for individuals, families, groups, agencies, businesses, communities, and systems (Taylor-Powell & Henert, 2008).
United Way (1996) made the distinction that outputs are about the program, while outcomes are about the participants. According to Taylor-Powell & Henert (2008), short-term outcomes are immediate and include changes in learning (i.e., awareness, knowledge, skills, attitudes, opinions, intent); medium-term outcomes are intermediate and include changes in action (i.e., behaviors, decision-making, actions); and long-term outcomes (often called impact) are the final outcomes which include changes in conditions (i.e., social, economic, civic, environmental). Assumptions are the beliefs about the program, the people involved, and how the program will work (Taylor-Powell & Henert, 2008). External factors include aspects that influence the way the program operates and are influenced by the program (Taylor-Powell & Henert, 2008).

It is important to note that a logic model is not a theory, is not reality, and is not an evaluation model or method (Taylor-Powell & Henert, 2008). Although not a theory, the logic model does allow one to display the theory of change for the program which is very helpful when using methods such as Chen’s theory-driven evaluation approach (Knowlton & Phillips, 2009). Knowlton and Phillips (2009) recommended using both a theory of change logic model and a separate program logic model. The theory of change model displays “what we will do” (strategies) and “what we will get” (results) (Knowlton & Phillips, 2009). The program logic model shows much more detail about the program and typically includes resources, activities, and outputs as part of the strategy and short-term, intermediate, and long-term outcomes as well as impact as part of the results (Knowlton & Phillips, 2009). Although not considered to be an evaluation model, a logic model does help one to determine what specifically to evaluate and what type of
evaluation approach would be most appropriate as shown in Figure 7. Providing the strongest evidence of program impact requires the evaluation of those longer-term outcomes.

![Logic Model and Common Types of Evaluation](image)

Figure 7: Logic Model and Common Types of Evaluation (Taylor-Powell & Henert, 2008)

There is risk of confusion as an evaluative outcome study does not always result in program outcomes as defined with regards to the logic model. An evaluative outcome study or the collection of outcome data can result in the reporting of only program outputs. Another discrepancy among terminology involves the term impact evaluation as
described in Figure 7. An impact evaluation is a form of outcome study and will be referred to as an evaluative outcome study for purposes of this investigation.

The logic model has essentially allowed program administrators to move beyond simply reporting program outputs when conducting evaluative outcome studies. Human service agencies or youth and family serving organizations typically report on how much money and staff/volunteers were involved; what the program did; and then how many participants were involved, how many hours were spent, and how many sessions were involved (United Way of America, 1996). This example of reporting involved inputs, activities, and outputs. These same organizations or agencies typically have not tracked what happened with participants after the services have been received (United Way of America, 1996). To move to a higher level of evidence of program impact, organizations must report on outcomes. The United Way of America (1996) defined outcomes as the benefits or changes for individuals or populations during or after participating in program activities. In other words, they are what participants know, think, or can do; or how they behave; or what their condition is, that is different following the program (United Way of America, 1996).

Documenting evidence of changes in social condition, a long-term outcome, can be a very difficult endeavor. For example, how can an organization document that an individual person who participated in an earlier program is now living a healthier lifestyle? Outcome indicators are used to document outcomes. Outcome indicators are specific items of data that can be tracked or measured to determine how well a program is achieving an outcome (United Way of America, 1996). In addition, outcome targets or
benchmarks can be established to gauge how well an outcome is being met (United Way of America, 1996).

The United Way of America (1996) has developed eight steps to follow for measuring program outcomes. The steps are as follows: (1) get ready; (2) choose the outcomes you want to measure; (3) specify indicators for your outcomes; (4) prepare to collect data on your indicators; (5) try out you outcome measurement system; (6) analyze and report your findings; (7) improve your outcome measurement system; and (8) improve your outcome measurement system. Each of these eight steps was designed to help program directors and managers better measure outcomes in programs dealing with human services.

Measuring Impact in Extension

The Kirkpatrick Model, Bennett’s Hierarchy, and the logic model are all tools that help Extension personnel conduct program evaluation and collect evidence to document impact. Keith Diem (2003) defined impact in Extension programming as “the positive difference we make in people’s lives as a result of the programs we conduct.” To make an impact, the results of an Extension program must ultimately change people’s attitudes or behavior, or benefit society in other ways (Diem, 1997). Another example of impact comes from private industry and the field of human resources. Brinkerhoff (2006) defined impact in organizational training and development as “taking the knowledge or skills learned from training and applying them to on-the-job behaviors which ultimately lead to worthwhile business results such as higher sales, reduced costs, or increased
retention of staff.” This may be translated to Bennett’s Hierarchy as experiencing KASA change as an immediate and direct result of the programming, application of the learning by demonstrating behavior/practice change, and impacting societal conditions by producing worthwhile end results.

Diem (2003) stated that evidence of more significant impact would be measured at the practice change and end results levels of Bennett’s Hierarchy. Diem (2003) admitted that lower levels of Bennett’s Hierarchy are important precursors, but not real evidence of significant impact. Even KASA change does not create much impact if behavior or practice change (application) does not follow and ultimately result in greater benefits to society. Impact must be documented by providing evidence of application. Clements (1999) stated that the ultimate goal of an Extension educator is to show evidence of the internalization of practices that Extension promotes. In order to document change at any level of impact, it is helpful to review the six stages of change: precontemplation, contemplation, information gathering, action, adoption, and internalization (Prochaska, Norcross, & DiClemente, 1994).

Behavior change is said to occur when one moves from one stage to the next (Prochaska et al., 1994). Therefore, evidence of behavior change and impact does not necessarily have to be complete internalization of a new method, but can simply be gathering information to potentially implement a new method. The literature associated with Kirkpatrick’s four levels reported that it is difficult to measure behavior change, and that larger corporations usually measure changes in behavior on a selective basis because of the expense and work required to create a custom design for each program (Dixon,
In addition, Clements (1999) reported that impact occurs when one moves from one stage to the next meaning it does not require complete mastery of the information provided by a program to have impact, but the impact on society might be much greater with complete mastery, application, and adoption.

What proportion of Extension personnel are currently conducting program evaluations that document evidence of higher level impact? The answer to this question is currently unknown. Previous research suggested that roughly one-third of the evaluation studies published in the *Journal of Extension* from 1998-2007 documented some evidence of effectiveness. Michael W. Duttweiler (2008) conducted a complete and thorough review of all articles published in the *Journal of Extension* during 1998-2007 to help determine if current evaluation practice is of value to the Cooperative Extension System. Program evaluation studies were defined using W. M. Trochim’s definition of “the systematic acquisition and assessment of information to provide useful feedback about some object”.

Duttweiler (2008) reported that an abstract must expressly reference the acquisition and use of programmatic information to be selected for further review. There were a total of 669 articles that met his criteria. Duttweiler (2008) assigned a “level of evaluation” based on the work of Francine Jacobs (1988) and her Five-Tiered Approach. The “levels of evaluation” and percentage of articles falling into that respective category were as follows: (1) needs assessment [~25%], (2) program documentation [<1%], (3) program fidelity (comparisons of program intent to actuality) [<1%], (4) program improvement [40%], and (5) evidence of effectiveness [35%]. Duttweiler (2008)
concluded that evaluation is both dynamic and influential in the Cooperative Extension System. However, measuring whether systems-wide impact has occurred would include only those which fall into the level (5) category (Bailey & Deen, 2002); therefore, Duttweiler’s study showed that 35% of all program evaluation articles published in the *Journal of Extension* over a ten-year period (1998-2007) were measuring some level of impact.

Duttweiler (2008) also characterized the state representation, program areas, geographic scope, and evaluation methods employed in the studies that met his criteria. For state representation, there were 48 states represented, two institutions accounted for 20% of the articles, and the top 10 institutions by number of publications accounted for approximately 50% of the articles (Duttweiler, 2008). The typical study was written for either youth development or agriculture and food systems program areas (~50%), had the dual purpose of outcome documentation and education process improvement, was statewide in scope (52%), and employed simple survey methodology (66%) (Duttweiler, 2008). In addition to the statewide studies, geographic scope included 32% local, 10% multicounty or in-state regions, and 6% multistate regions (Duttweiler, 2008).

Measuring more significant impact is typically accomplished by documenting evidence of behavior/practice change (medium-term outcomes) and end results (long-term outcomes) (Diem, 2003). Programs are often evaluated by surveying participants regarding their level of satisfaction with the teaching (instructor and/or content). Another example would be a form which requests that participants rate their level of agreement regarding statements about the quality of the teaching or instruction. This is typically
accomplished by using a questionnaire that is administered at the end of the program. Various sources refer to these types of evaluations as using “smile sheets” or “happy sheets”. These evaluations can provide information that is very useful to program administrators; however, the evaluation instruments are not likely designed for measuring higher level impact documented by behavior change (medium-term outcome) or end results (long-term outcome). This type of evaluation would typically be used to measure participants’ reactions. This would be very similar to the student evaluation of instruction (SEI) utilized for evaluating courses in higher education. Ohio State University Extension has developed the Evaluation of Effective Extension Teaching (EEET) instrument to measure teaching effectiveness. The EEET provides a way for Extension personnel to measure participants’ reaction to the teaching (instructor and/or content); however, much more is needed to document evidence of higher level or more significant impact (i.e., behavior change and end results).

**Reach and Richness**

Evans and Wurster’s (2000) concepts of “reach and richness” have not been previously investigated with regards to program evaluation. David King and Michael Boehlje (2000) were the first to refer to these concepts when discussing the relevance of Cooperative Extension. Their point was that Extension must be concerned with ‘reach’ and remain relevant or private industry will attempt to fill the void. Evans and Wurster (2000) have written about this continuum regarding physical modes of delivery of information and the inability to improve one aspect without sacrificing the other (Figure 53).
However, they have determined that new technology now allows for organizations to do both when delivering information. Digital networks allow for a greater number of people to exchange very rich information without the traditional constraints of cost and proximity. In order for the continuum to be broken, there must be a displacement of connectivity and the adoption of common standards (Evans & Wurster, 2000).

![The Traditional Economics of Information](image)

Figure 8: Traditional Economics of Information (Evans & Wurster, 1997) (Richness also includes reliability, security, and currency in addition to bandwidth, customization, and interactivity.)

When information is carried by a physical thing, it goes wherever that thing goes (Evans & Wurster, 1997). When people are connected electronically, information can then travel by itself (Evans & Wurster, 1997). Electronic connectivity allows information to become unbundled from its physical carrier (Evans & Wurster, 1997). Reach is easily
understood as the number of people or groups that participate or are touched by receiving information.

Richness is more complex and more difficult to fully understand. Evans and Wurster (2000) have defined richness by six aspects of the information itself. The first is bandwidth, or the amount of information that can be moved from sender to receiver in a given time (e.g., stock quotes are narrowband and a feature film is broadband). The second is the degree to which the information can be customized (e.g., advertisement on television is far less customized than a personal sales pitch, but reaches far more people). The third is interactivity (e.g., dialogue is possible for a small group, but to reach millions of people the message must be a monologue). The fourth is reliability (e.g., information is reliable when exchanged among a small group of trusted individuals, but is not when it is circulating among a large group of strangers). The fifth is security (e.g., managers share highly sensitive business information only in closed-door meetings, but they will disseminate less sensitive information to a wider audience). The sixth and final aspect is currency (e.g., on Wall Street, where seconds count, a few market makers have instantaneous quotes, a larger group of financial institutions receives quotes with three- to fifteen-minute delays, and most retail investors receive quotes with at least a 15-minute delay).

Traditionally, the communication of rich information has required proximity and dedicated channels whose costs or physical constraints have limited the size of the audience to which the information could be sent; conversely, the communication of information to a larger audience has required compromises in bandwidth, customization,
interactivity, reliability, security, and currency (Evans & Wurster, 2000). However, this trade-off changes in a digital world.

The future may dictate that Extension must exclusively provide information with tremendous richness that reaches a tremendous amount of people in order to remain relevant. However, for the time being, these terms do an excellent job of explaining the types of delivery methods traditionally and currently used in Extension programming. For example, an identified need could be addressed by writing a fact sheet that is sent out to an extraordinarily large amount of people. This is high reach, but low richness because fact sheets are typically not designed for the delivery of a large amount of very in-depth information. Alternatively, a short course could be held in which a few people are exposed to tremendously detailed information. This would result in high richness, but very low reach because only a few people can actually enroll in the short course and receive the in-depth instruction necessary. Which of these choices is the correct one to make? The answer should be the one that results in the greatest impact. This study will attempt to determine if there are any relationships between these terms and documenting evidence of impact.

**Geographic Scope**

For the purposes of this study, geographic scope will consist of the categories of county (local), multicounty regional, statewide, multistate regional, or national. Each U.S. state and territory has a state office located at its land-grant university and a network of local or regional offices (USDA NIFA, 2010). The number of local Extension offices
has declined over the years, and some county offices have consolidated into regional centers; however, there are still approximately 2,900 Extension offices located nationwide (USDA NIFA, 2010). Extension expertise has traditionally prided itself on meeting public needs at the local level. Historically, programming was targeted towards local communities. Modern technology and transportation have allowed programming to be targeted towards much larger audiences and clientele located across state lines. Some programs are now created and targeted for the entire nation.

The geographic scope of programming is important and may have implications regarding future changes in structure and arrangement of statewide Extension systems. Most states have traditionally used a county cluster model in which Extension Educators were located or housed within a single county, but often worked in the neighboring counties (Ahmed & Morse, 2010). Cooperative Extension in the state of Minnesota changed this paradigm by adopting a regional/county model in which Extension Educators worked out of a regional center responsible for several counties within the region (Ahmed & Morse, 2010). By determining any potential impact differences as a result of geographic scope, states may be able to adjust their arrangement to confront financial issues.

Summary

To summarize, the Cooperative Extension System is very complex with tremendous diversity of information (content) and a variety of delivery methods (process). Impact is likely greater when there is both higher content and higher process.
Simply providing a service is not going to produce the same amount of impact as providing real educational training. Program evaluation is the method used to collect and document evidence of impact. There are different levels of evidence of impact ranging from simply measuring the reactions of participants to measuring the economic impact on the nation’s economy. Models help one to conduct program evaluation and use evaluative outcome studies to collect specific levels of evidence of impact and properly describe that evidence.

The first visual display model developed for this purpose was Kirkpatrick’s model for evaluating training programs. The next model to be developed was Bennett’s Hierarchy created for the evaluation of Cooperative Extension programming. Finally, the logic model was created and widely adopted because of its wide range of uses in addition to evaluation. The purpose of these models is to help collect and describe evidence and to document program impact. Impact is important for the accountability of programs. Measuring and documenting impact helps to ensure continued and additional funding. In addition, it helps an organization to improve existing programs and create new programs.

Michael Duttweiler examined evaluation studies published in the Journal of Extension from 1998-2007 to determine the value of program evaluation in Extension. He found that 35% of all studies dealing with evaluation results were showing some evidence of effectiveness. It is still unknown as to the level of evidence of impact measured by program evaluation in Extension. In addition, it is unknown how the measurement of impact has progressed since 1965 when the Journal of Extension and Kirkpatrick’s model were still very new and Bennett’s Hierarchy and the logic model
were yet to be created. Finally, the reach, richness, and geographic scope of
programming have an effect on the overall value of impact. It is less likely that
programming with very low reach will make significant changes in conditions for society
unless that programming also has tremendous richness.
CHAPTER 3

METHODS

Research Design

The purpose of this study was to explore and describe the level of evidence of impact collected through program evaluation (outcome studies) by the Cooperative Extension System as published in the *Journal of Extension*; and to examine any potential relationships between the documentation of impact with reach, richness, and geographic scope. Impact was operationally defined as program effectiveness determined by the highest level of evidence collected based on the components of both Bennett’s Hierarchy (e.g., inputs, activities, people involvement, reactions, KASA change, practice change, and end results) and the University of Wisconsin-Extension logic model format (e.g., situation/priorities, inputs, outputs-activities, outputs-participation, short-term outcomes, medium-term outcomes, and long-term outcomes).

Abstracts for all articles published in the *Journal of Extension* under the headings of ‘Feature Articles’ and ‘Research in Brief’ during five separate five-year periods (1965-69, 1975-79, 1985-89, 1995-99, and 2005-09) were assessed to identify those that reported on program evaluation and were considered outcome studies. Articles that met the appropriate criteria (evaluation of an existing program) were selected and examined further to complete the instrument.
This study utilized a form of quantitative content analysis where data was collected and coded using qualitative methods and analyzed using quantitative methods. Riffe, Lacy, and Fico (1998) defined quantitative content analysis as the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, in order to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption. The articles were reviewed qualitatively and the instrument was completed by making a professional judgment and coding the appropriate choice for each item based on predetermined criteria contained in a rubric. Quantitative data analysis procedures were used to interpret the coded data.

Population and Sampling

The target population for this study consisted of all published reports pertaining to the evaluation of Extension programs nationwide during five separate five-year periods (1965-69, 1975-79, 1985-89, 1995-99, and 2005-09). The accessible population consisted of those articles pertaining to program evaluation (outcome studies) and published in the selected categories of the *Journal of Extension* during these time frames. Sampling procedures were not needed as this study utilized a census of the accessible data during the time frames investigated.

Articles pertaining to program evaluation were selected by adapting similar procedures previously developed and utilized by Michael W. Duttweiler (2008). The
operational definition used for evaluation was any work that involved “the systematic acquisition and assessment of information to provide useful feedback about some object” (Trochim, 2006). Just like Duttweiler’s (2008) investigation, articles selected had to expressly reference the acquisition and use of programmatic information. Consequently, articles not selected were those which were commentaries on evaluation needs or methods without reference to supporting evaluation data, applied research reports in content areas that might inform Extension curricula but that did not specifically address educational processes, and descriptions of evaluation resources or methods without reference to application data (Duttweiler, 2008). For this study, more restrictive selection procedures were developed to ensure that only outcome studies were selected.

The Journal of Extension was chosen for purposes of this study because it was reported to be the official refereed journal of the Cooperative Extension System (Extension Journal, Inc., 2010). According to Extension Journal, Inc. (2010), the journal was written, reviewed, and edited by Extension professionals for the purpose of sharing successful educational applications, original and applied research findings, scholarly opinions, educational resources, and challenges on issues of critical importance to Extension educators. Originally known as the Journal of Cooperative Extension, the journal was first published during the spring of 1963. The name was officially changed to the Journal of Extension in 1969. The first piece published in 1963 under ‘About this Issue’ stated that the journal “is dedicated to the professional growth and advancement of those who have chosen Extension education as a career and to other leaders in adult education and community development” (Ferguson & Carter, 1963). The journal was
published in print form during its first 30 years of existence, but has been published entirely online and distributed electronically since 1994. Those issues originally produced in print form have been archived electronically in Portable Document Format (PDF) on the journal’s web site (http://www.joe.org).

The *Journal of Extension* contained approximately 2,784 articles and/or pieces presented in the 130 issues published during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. Many of the writing pieces contained under certain headings (e.g., ‘Book Review’ and ‘Letters to the Editor’) could not be described as articles; therefore, these will be referred to as pieces. The journal was published four times per year during 1965-69 and 1985-89 and six times per year during 1975-79, 1995-99, & 2005-09. Each of the articles/pieces were published under the following categories/headers with the total number listed in parentheses: Feature Articles (730), Research in Brief (516), Ideas at Work (333), Tools of the Trade (367), Editor’s Page (105), Commentary (67), Idea Corner (53), Abstracts (384), Book Review (78), Point of View (61), To the Point (15), Forum (39), Futures (18), Letter to the Editor (11), About this Issue (4), and Bibliography (3).

This investigation utilized only those articles contained under the categories of ‘Feature Articles’ and ‘Research in Brief’ (N = 1,246). ‘Feature Articles’ have a maximum of 3,000 words and discuss concepts and research findings of particular interest and significance to U.S. Extension professionals and to U.S. Extension's knowledge base, methodology, effective practice, and organization (Extension Journal, Inc., 2010). ‘Research in Brief’ have a maximum of 2,000 words and summarizes
research results of importance to U.S. Extension professionals (Extension Journal, Inc., 2010). According to the Extension Journal, Inc. (2010), the difference is that ‘Feature Articles’ are broader in scope and implication focusing on the implications of the data or concepts for a wide audience of U.S. Extension professionals whereas ‘Research in Brief’ is more specific and localized focusing more on the data itself, and the methods used to gather it.

Articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’ are typically the only articles which contain enough depth and detail necessary to complete the instrument utilized for this study. Furthermore, articles contained in these two sections were peer reviewed by multiple (three) reviewers whereas articles published under ‘Ideas at Work’ were peer reviewed by a single reviewer. The remaining sections could not be considered as peer reviewed as they were reviewed solely by the editor (Extension Journal, Inc., 2010).

It is important to note that articles published under the major headings of ‘Ideas at Work’ and ‘Tools of the Trade’ were not reviewed as part of this study; although, these sections may contain some articles which report outcome data from program evaluation and the documentation of evidence of impact. The ‘Ideas at Work’ section contains articles with a maximum of 1,000 words that describe novel ideas, innovative programs, and the new methods of interest to U.S. Extension professionals (Extension Journal, Inc., 2010). Articles in the ‘Tools of the Trade’ section contains a maximum of 1,000 words and reports on specific materials, books, techniques, and technologies useful to U.S. Extension professionals (Extension Journal, Inc., 2010). According to the Extension
Journal, Inc. (2010), these two sections are very similar, but ‘Tools of the Trade’ focuses on a thing rather than an idea. In summary, only articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’ were chosen for this study because these categories (1) typically contained lengthier articles with enough depth and detail to complete the instrument; (2) likely contained most of the pertinent articles reporting program evaluation results/findings; and (3) were peer reviewed by multiple reviewers.

Instrumentation

The instrumentation procedures utilized for this study would be best described as using a researcher-completed observation form or tally sheet (Fraenkel & Wallen, 2006). The instrument consisted of a Microsoft® Office Word® 2007 document (see Appendix B) to use as a questionnaire and a Microsoft® Office Excel® 2007 spreadsheet (see Appendix C) to use for entering data. In addition, a rubric (see Appendix A) was created to help ensure that data was collected accurately and consistently.

The instrument can be described as having three sections (organizational information, Stage I, Stage II). The first portion of the instrument consisted of organizational data (similar to demographic data) which was pre-entered into the spreadsheet prior to article selection and formal data collection. This section includes the publication year, volume, number, lead author’s last name, and the category of JOE in which the article was listed. The second portion involved question 1 on the questionnaire which was completed during Stage I by screening the abstracts/articles to determine if the article should be selected for formal data collection. Question 1 simply asks if the article
should be selected for use in this study. The third portion consisted of the actual data collection (Stage II) which involves coding the data and completing the questionnaire.

Stage II included the remainder of the questionnaire (questions 2 – 12) and was completed by reading the articles selected during Stage I. Question 2 requires entering the lead author’s home state at the time of publication (nominal data) and is taken directly from the author’s title or e-mail address. This information was used to determine the coverage or portion of the nation represented in this study. Question 3 requests the occupation of the lead author (nominal data) taken directly from their title and entered as: 1 (Extension Professional – Agent/Educator), 2 (Extension Profession – Specialist), 3 (Extension Profession – Director, Leader, Head, Administrator), 4 (Extension Staff – Extension Associate; Program Assistant/ Coordinator/ Manager), 5 (University Faculty – Assistant Professor/ Associate Professor/ Professor), 6 (University Student – Graduate Assistant/ Associate), 7 (other), or 8 (unknown). The occupation of the primary author was used to determine who typically published program evaluation results in the *Journal of Extension*.

Question 4 was the primary program areas (nominal data) and consisted of the four traditional program areas as defined by Seevers, Graham, and Conklin (2007): 1 (Agriculture and Natural Resources), 2 (4-H Youth Development), 3 (Family and Consumer Sciences), 4 (Community and Economic Development), and 5 (combination of program areas). Question 5 involved the object of the evaluation in which were as follows: 1 (Individual Programs for Extension Clientele/Volunteers [evaluatees such as: participants or recipients]), 2 (Educational Tools for Extension Clientele/Volunteers
evaluands such as: newsletters, bulletins, web sites, television or radio programs], 3 (Individual Programs for Extension Personnel [e.g., in-service / training programs], 4 (Program Area [i.e., Agr & NR, 4-H, FCS, CED]), and 5 (U.S./State Cooperative Extension System). Question 6 involved the level of evidence collected according to Bennett’s Hierarchy which ranged from 1 to 7 (e.g., 1 = inputs, 2 = activities, 3 = people involvement, 4 = reactions, 5 = KASA change, 6 = practice change, 7 = end results). Questions 7 involved the level of evidence collected according to the University of Wisconsin-Extension logic model format also ranged from 1 to 7 (e.g., 1 = situation/priorities, 2 = inputs, 3 = outputs-activities, 4 = outputs-participation, 5 = short-term outcomes, 6 = medium-term outcomes, 7 = long-term outcomes). Question 8 concerned the primary evaluation method and was reported using Diem’s (2003) three essential ways to evaluate impact: 1 (ask them-survey), 2 (test them-simple experimental designs), or 3 (observe them-observations).

Question 9 adapted the primary data collection methods also identified from Diem’s (2003) examples of basic methods and tools used to measure impact: 1 (written questionnaires), 2 (longitudinal studies), 3 (interviews/testimonials/case studies/focus groups), 4 (pre-test/post-test), 5 (post-test with control group comparison [i.e., both true- and quasi- experimental]), 6 (direct observation), 7 (reviewing information from other sources), 8 (online survey), 9 (retrospective pretest / post-then-pre), and 10 (other). Questions 10 and 11 involved reach and richness and both were recorded separately as either 1 (low) or 2 (high). Question 12 involved geographic scope and was entered as
either 1 (county [local]), 2 (multicounty regional), 3 (statewide), 4 (multistate regional), or 5 (national).

The level of evidence of impact was quantified by coding ordinal data according to both Bennett’s Hierarchy (ranging from 1 to 7) and the UWEX logic model format (ranging from 1 to 7). Ratings of high or low for reach and richness were collected as ordinal data and were adapted from the definitions provided by Evans and Wurster (1997; 2000) and King and Boehlje (2000). Geographic scope was also collected as ordinal data.

Reach is the amount of individuals you can touch and influence (with information). The criteria for high reach programs were those that generally (1) were created and targeted to reach a larger audience, (2) may not require any formal registration or attendance, and/or (3) may not contain any evidence or record of participation by any single individual. An example of high reach programs would be those providing information through mass media. Criteria for programs rated as low reach were the exact opposite with programs that generally (1) were created and targeted to reach a smaller and more well-defined audience, (2) may not require formal registration and/or attendance, and/or (3) may not contain evidence of participation and attendance.

Richness is the detail and depth of the interaction (with information). For example, programs that require participants to attend or participate multiple times allowing them the opportunity to practice and apply their new knowledge and receive feedback and more individualized instruction throughout the program’s duration are
typically considered as having high richness. Generally, one shot programs have low richness. Criteria for high richness generally consisted of (1) high bandwidth in which very in-depth information was moved from sender to receiver, (2) high customization in which the program was specifically tailored to a specific audience, (3) high interactivity in which there was a great deal of exchange between program providers (e.g., instructors, administrators) and participants, (4) high reliability in which participants tended to trust program providers, (5) high security in which sensitive information could be shared, and/or (6) high currency in which information exchange (feedback) was immediate.

Criteria for low richness were considered the exact opposite with programs that generally had (1) low bandwidth in which very general information was provided, (2) low customization in which the program was not specifically targeted to a specific audience, (3) low interactivity in which there was little to no exchange between program providers (e.g., instructors, administrators) and participants, (4) low reliability with less trust as participants and program providers were likely strangers, (5) low security in which sensitive information was unlikely to be shared, and/or (6) low currency in which information exchange (feedback) was slower.

Reach is usually negatively correlated with richness unless there was a digital network or electronic means utilized for the delivery of the information (e.g., web sites, CD ROM, television, radio). This is because when information is embedded in physical modes of delivery there is a universal trade-off between reach and richness (Evans & Wurster, 1997; 2000). However, more complex Extension programming (especially
when the evaluation object is program area or U.S./State CES) can and often does both high reach and high richness and should be coded accordingly (positively correlated).

Completion of the full instrument required the reading and reviewing of abstracts/articles which was considered to be content analysis using qualitative data collection procedures. Portions of this data recording process were somewhat subjective as the researcher had to make a judgment during the screening process as to whether the abstract pertained to the evaluation of existing programming. Furthermore, the researcher had to make some professional judgments in determining the primary program area(s); identifying the object of the evaluation; recognizing the level of evidence of impact collected according to Bennett’s Hierarchy and the UWEX logic model; identifying the evaluation and data collection method(s); and appropriately coding reach, richness, and geographic scope. Due to the subjective nature of the instrument, the instrument was initially field tested and pilot tested by a former Extension faculty administrator. Furthermore, estimates of inter-coder agreement (inter-rater reliability) and intra-coder agreement (intra-rater reliability) were determined.
**Validity and Reliability.** Both the field and pilot tests were completed simultaneously by a former Extension administrator trained and familiar with the purpose and methods of this study. The field test was used to determine that the instrument and rubric were suitable and the instructions and procedures developed by the researcher were clear and executable. In addition to the feedback received from the dissertation committee, the field test helped to establish face and content validity of the instrument.

The pilot test was used to determine the consistency or repeatability of the instrument and give an initial estimate of reliability. The pilot test consisted of both the researcher and the former Extension administrator independently completing the instrument for the February, 2005 (volume 43, number 1) issue. The former administrator read chapters 1-3 of this document and was briefly trained by the researcher prior to data collection. The results reported by the researcher were compared to the results obtained by the expert and it was determined that there was 100% agreement for Stage I and 80% agreement for Stage II.
**Inter-rater Reliability.** Inter-rater reliability measures reproducibility by determining if coding schemes lead to the same text being coded in the same category by different people (Stemler, 2001). To test the inter-coder agreement (inter-rater reliability) of the instrument, two raters were selected to review randomly selected JOE issues. Rater A was the same former Extension Administrator who assisted with the initial field and pilot testing. Rater B was a doctorate student with a strong Agricultural Education background who was specializing in teacher education. Rater A had a great deal of experience with Cooperative Extension while Rater B had very limited familiarity. The reliability process was also divided into Stage I and Stage II. Each of the raters received a copy of the proposal (chapters 1-3), questionnaire, rubric, and an abridged version of the spreadsheet containing only the articles from the randomly sampled issues. The raters initially read chapters 1-3 of this document to gain some background information on program evaluation and the documentation of evidence of impact as well as to gain familiarity with the purpose and methods of this study. The raters were then briefly trained by the researcher as to how to screen abstracts and complete the instrument for selected articles. The training also involved practicing the process of selecting articles and collecting data from articles not part of the sample.

A random sample of 13 (13 / 130 = 10%) individual issues was selected for completion by the two raters. In order to ensure that the sample contained an adequate distribution of the published articles, a stratified random sampling process was used so that two issues were selected from both the 1960s and 1970s; and three issues were selected from each of the more recent decades (1980s, 1990s, and 2000s).
For Stage I, the raters screened all abstracts/articles within the randomly sampled issues under the applicable headings of ‘Feature Articles’ and ‘Research in Brief’. The randomly sampled issues ultimately contained 124 (124 / 1,246 = 9.95%) of the total articles reviewed for this study. The results from each of the raters were compared to that obtained by the researcher and inter-coder agreement was calculated for Stage I using Cohen’s Kappa (inter-rater reliability coefficient), percent agreement, and phi (correlation coefficient).

The interpretation of Kappa values was based on the findings of Landis (1977) and interpreted as follows: less than chance agreement (< 0); slight agreement (0.01-0.20); fair agreement (0.21-0.40); moderate agreement (0.41-0.60); substantial agreement (0.61-0.80); and almost perfect agreement (0.81-0.99). According to Viera (2005), this is a commonly cited scale used for interpreting Kappa. Correlation coefficients were also used to determine the magnitude and direction of the association. The magnitude was interpreted using Bartz’s (1999) terminology as follows: very high (0.80 or higher), strong (0.60 to 0.80), moderate (0.40 to 0.60), low (0.20 to 0.40), and very low (0.20 or lower).

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selected</td>
<td>Not Selected</td>
<td></td>
</tr>
<tr>
<td>Researcher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selected</td>
<td>36</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>Not Selected</td>
<td>8</td>
<td>73</td>
<td>81</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>80</td>
<td>124</td>
</tr>
</tbody>
</table>

Table 1: Stage I Cross Tabulation between Researcher and Rater A (Kappa = 0.734; percent agreement = 87.9; phi = 0.735)
As shown in Table 1, there was substantial agreement and a strong association between the researcher and Rater A for Stage I as each agreed 87.9% of the time. There was moderate agreement and a moderate association between the researcher and Rater B as each agreed 72.6% of the time (Table 2).

For Stage II, the raters received another abridged version of the spreadsheet containing only the articles actually selected by the researcher during Stage I within those randomly sampled issues. The abridged spreadsheet ultimately contained 43 (43 / 302 = 14.2%) of the total articles selected during Stage I for the entire study. Inter-coder agreement using percent agreement values were calculated for Stage II to compare the raters’ results with the researcher’s results. In addition, Spearman rank-order correlation coefficients were calculated for ordinal data collected for Level-Bennett’s Hierarchy and Level-UWEX Logic Model. The values were interpreted based on the same criteria previously used for Stage I.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Rater A Percent Agreement</th>
<th>Rater A Spearman’s rho</th>
<th>Rater B Percent Agreement</th>
<th>Rater B Spearman’s rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of Lead Author</td>
<td>95.4</td>
<td></td>
<td>95.3</td>
<td></td>
</tr>
<tr>
<td>Occupation of Lead Author</td>
<td>83.7</td>
<td></td>
<td>79.1</td>
<td></td>
</tr>
<tr>
<td>Primary Program Area</td>
<td>97.7</td>
<td></td>
<td>93.0</td>
<td></td>
</tr>
<tr>
<td>Object of the Evaluation</td>
<td>86.1</td>
<td></td>
<td>62.8</td>
<td></td>
</tr>
<tr>
<td>Level-Bennett’s Hierarchy*</td>
<td>86.1</td>
<td>0.617</td>
<td>65.1</td>
<td>0.681</td>
</tr>
<tr>
<td>Level-UWEX Logic Model*</td>
<td>81.4</td>
<td>0.636</td>
<td>67.4</td>
<td>0.617</td>
</tr>
<tr>
<td>Primary Evaluation Method</td>
<td>75.6</td>
<td></td>
<td>77.9</td>
<td></td>
</tr>
<tr>
<td>Primary Data Collection Method</td>
<td>72.1</td>
<td></td>
<td>76.7</td>
<td></td>
</tr>
<tr>
<td>Level of Reach</td>
<td>76.7</td>
<td></td>
<td>55.8</td>
<td></td>
</tr>
<tr>
<td>Level of Richness</td>
<td>81.4</td>
<td></td>
<td>77.4</td>
<td></td>
</tr>
<tr>
<td>Geographic Scope</td>
<td>67.4</td>
<td></td>
<td>41.9</td>
<td></td>
</tr>
<tr>
<td>Overall Agreement</td>
<td>84.0</td>
<td></td>
<td>76.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Stage II Percent Agreement between Researcher and Raters by Variable

*Multichotomous ordinal data considered as agreement when within +/- 1

As shown in Table 3, overall agreement was 84.0% between the researcher and Rater A and 76.6% between the researcher and Rater B. There was a strong association for both Level-Bennett’s Hierarchy and Level-UWEX Logic Model between the researcher and both raters. The only major concern was that the percent agreement for level of reach with Rater B was 55.8%. The values for Rater B were possibly lower because Rater B had less familiarity and understanding of Cooperative Extension as well as program evaluation.
Intra-rater Reliability. Intra-rater reliability measures stability by determining if the same coder can get the same results try after try (Stemler, 2001). The researcher collected additional data to establish intra-rater reliability after the formal data collection process was entirely completed. The length of time between the initial collection and the second collection varied among the randomly selected issues, but ranged from around two months to about two weeks.

The intra-coder agreement (intra-rater reliability) was established by taking a random sample of seven JOE issues (7 / 130 = 5.39%) and having the researcher repeat both the article selection process (Stage I) and data collection process (Stage II) for these individual issues. In order to ensure that the sample contained an adequate distribution of the published articles, a stratified random sampling process was used so that at least one issue was selected from each decade. The random sample ultimately contained 14 (14 / 302 = 4.6%) of the total articles selected during Stage I for the entire study. The new results were then compared to the original results and Cohen’s Kappa, percent agreement, and the phi coefficient were calculated for Stage I. Percent agreement values were calculated for Stage II.

<table>
<thead>
<tr>
<th>Researcher’s Second</th>
<th>Selected</th>
<th>Not Selected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher’s First</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selected</td>
<td>14</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Not Selected</td>
<td>0</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>51</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 4: Stage I Cross Tabulation between Researcher’s First and Second Selections (Kappa = 1.0; percent agreement = 100.0; phi = 1.0)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of Lead Author</td>
<td>100.0</td>
</tr>
<tr>
<td>Occupation of Lead Author</td>
<td>92.9</td>
</tr>
<tr>
<td>Primary Program Area</td>
<td>94.6</td>
</tr>
<tr>
<td>Object of the Evaluation</td>
<td>100.0</td>
</tr>
<tr>
<td>Level-Bennett’s Hierarchy*</td>
<td>100.0</td>
</tr>
<tr>
<td>Level-UWEX Logic Model*</td>
<td>100.0</td>
</tr>
<tr>
<td>Primary Evaluation Method</td>
<td>92.9</td>
</tr>
<tr>
<td>Primary Data Collection Method</td>
<td>92.9</td>
</tr>
<tr>
<td>Level of Reach</td>
<td>78.6</td>
</tr>
<tr>
<td>Level of Richness</td>
<td>78.6</td>
</tr>
<tr>
<td>Geographic Scope</td>
<td>85.7</td>
</tr>
<tr>
<td>Overall Agreement</td>
<td>92.9</td>
</tr>
</tbody>
</table>

Table 5: Stage II Percent Agreement between Researcher’s First and Second Collections by Variable
* Multichotomous ordinal data considered as agreement when within +/- 1

The findings for intra-rater reliability were higher than those found for inter-rater reliability. For Stage I, there was perfect agreement (Table 4). For Stage II, overall percent agreement was 92.9% (Table 5). Intra-rater reliability was considered important because the selection of articles and collection of data which were analyzed to address the research objectives of this study were collected exclusively by the researcher.

Data Collection

All data utilized for this study were manually collected and entered into the instrument by the researcher. Prior to beginning the formal data collection process, the researcher conducted numerous field tests with volumes published during various decades to help work out any problems and inadequacies of the instrumentation procedures and to create the rubric. This was considered to be a normal and required part of the process when using content analysis and coding data.
Initially, organizational data was pre-entered into the spreadsheet to help locate articles and to enter data into the correct rows and columns. Once the instrumentation process was determined to be ready, the formal data collection consisted of a two-stage process. Stage I involved screening abstracts and scanning articles to select the articles that were appropriate for this study. The Spring, 1965 (volume 3, number 1) issue of the *Journal of Extension* was used as the starting point for the formal data collection process. Each of the articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’ were screened by reading the abstract online. This process was then repeated for the 129 remaining issues. In order to help ensure consistency, Stage I was completed in its entirety prior to beginning Stage II. Stage II was also completed by accessing and reading the articles online. Stage II involved reading the selected articles, coding the text, and completing the remainder of the instrument.

Articles were selected during stage I by screening the abstract and answering five questions in which all must be answered as “yes”.

1. Does the article involve the systematic acquisition and assessment of information to provide useful feedback about some object?
2. Does the article involve the acquisition and use of programmatic information?
3. Can the article be considered as a program evaluation meaning was the purpose to determine the value (worth or merit) of something? (How good is something?)
4. Can the type of program evaluation be considered as an outcome study in that an intervention has already occurred and outcome data was collected?
5. Can an object of the evaluation be identified and does it belong to the U.S. Cooperative Extension System (CES)?

The evaluation object was defined as whatever is being evaluated. Identifying and thinking about the object of the evaluation was necessary to determine if an article should be selected for this study. Some articles claimed to be program evaluations or evaluative studies while others were simply research studies (i.e., exploratory/descriptive, explain/predict, experimental). Instead of determining if an article was research or evaluation, a judgment was required as to whether the results could be considered as evaluative outcome data.

To be selected for the purposes of this study, the object of the evaluation had to be some form of educational programming; either an individual program, an educational tool, or a program system. It was required that the educational programming was a part of the U.S. Cooperative Extension System meaning that it was essential that CES conducted the intervention (or programming); however, partnering with other organizations or entities was considered acceptable. When the evaluation object was an educational tool, CES must have created, modified, or utilized the tool for Extension programming. Although Extension personnel occasionally evaluate non-Extension programs or objects, these types of articles were not selected for purposes of this study as CES did not conduct the intervention.

Five additional criteria were created to help eliminate (or not select) an article while screening the abstract.

1. No data or information was collected.
2. The data or information cannot be considered as programmatic information.

3. The article is clearly a research study in which the results could not be used for evaluation (determining the value [worth or merit] of an object).

4. The article was not reporting the results from an outcome study. (It was a needs assessment or process study [monitoring study]).

5. The object is not considered to be a part of the U.S. Cooperative Extension System.

Articles were not selected, if the object was not considered to be part of CES meaning that for programs, there was no intervention by CES. For educational tools, they were not created, modified, or utilized by CES.

Articles regarding the assessment of needs were not selected if the established criteria were not satisfied. Needs assessments were only selected if the assessment was somehow already incorporated into an outcome evaluation (determining value) being reported in the article. For example, a current program assesses what else is needed while collecting evaluative information regarding what has already been delivered. Process or monitoring studies were also not selected if the previous criteria were not satisfied. Process studies were only selected if they included outcome data regarding an individual program. Articles were eliminated in Stage II after selection (abstract screening process) when it was eventually determined that the respective article did not meet the established criteria.

Stage II consisted of the actual data collection, and additional criteria were created and printed in the rubric to assist with coding and entering the appropriate
response. At the conclusion of the data collection process, all data was transferred from Microsoft® Office Excel® to IBM® SPSS® Statistics 18.0 for analysis.

Data Analysis

Data was analyzed to satisfy each objective of the study using IBM® SPSS® Statistics 18.0. Frequencies and percentages were calculated to report descriptive information for the articles that were reviewed for this study. In addition, the nonparametric Kruskal-Wallis test was used to calculate mean ranks and determine the differences between objects of the evaluation. The research objectives and data analysis procedures utilized to satisfy each objective were as follows:

1. To determine the number of program evaluations with outcome data published in the Journal of Extension during the years investigated and the program areas represented in these evaluation studies.

   Frequencies and percentages were reported for ratio data.

2. To review a subset of articles published in the Journal of Extension from 1965-2009 in order to determine the number and percentage of evaluation studies that reached each level of evidence of impact according to both Bennett’s Hierarchy and the logic model.
Frequencies and percentages were reported for ratio data. In addition, percentiles and medians as nonparametric descriptive data for the ordinal Level-Bennett’s Hierarchy and Level-UWEX Logic Model were reported.

3. **To investigate and identify any trends over time.**

   Frequencies and percentages were reported for each level of evidence during each decade (ordinal data). Frequencies and percentages were also reported for groupings of level of evidence. In addition, the Kruskal-Wallis test was used to calculate mean ranks and determine the differences between decades.

4. **To identify and quantify the primary evaluation and data collection methods used to document evidence of impact.**

   Frequencies and percentages were reported for nominal data.

5. **To determine if the program being evaluated was designed for reach (number of people interacted with) and/or richness (detail and depth of interaction).**

   Frequencies and percentages were reported for ordinal data as well as a cross tabulation between reach and richness.

6. **To determine the geographic scope (county [local], multicounty regional, statewide, multistate regional, or national) of the programs evaluated.**

   Frequencies and percentages were reported for ordinal data.
7. To explore the relationship between reach/richness, geographic scope, and the level of evaluation according to both Bennett’s Hierarchy and the logic model.

Spearman rank-order correlation coefficients were reported for each ordinal variable.
CHAPTER 4

RESULTS

The purpose of this study was to explore and describe the level of evidence of impact collected through program evaluation (outcome studies) by the Cooperative Extension System as published in the *Journal of Extension*; and to examine any potential relationships between the documentation of impact with reach, richness, and geographic scope. Impact was operationally defined as program effectiveness determined by the highest level of evidence collected based on the components of both Bennett’s Hierarchy (i.e., 1 = inputs, 2 = activities, 3 = people involvement, 4 = reactions, 5 = KASA change, 6 = practice change, 7 = end results) and the UWEX logic model format (i.e., 1 = situation/priorities, 2 = inputs, 3 = outputs-activities, 4 = outputs-participation, 5 = short-term outcomes, 6 = medium-term outcomes, 7 = long-term outcomes). Articles reviewed in the *Journal of Extension* were those listed under the headings of ‘Feature Articles’ and ‘Research in Brief’ during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. To achieve this stated purpose, a set of research objectives was created. This chapter presents descriptive information regarding the articles that were investigated as well as the findings pertaining to each of the seven research objectives which ultimately guided this study.
The formal data collection consisted of a two-stage process. Stage I (selection of articles) involved screening abstracts and scanning articles to select those that were appropriate based on the criteria established for this study. Stage II (collection of data) involved reading the selected articles, coding the text, and completing the remainder of the instrument.

There were a total of 1,246 articles investigated for purposes of this study. There were 730 categorized as ‘Feature’ articles and 516 categorized as ‘Research in Brief’ articles. The lead authors of the articles selected during Stage I (302) represented 47 different states as well as Canada and Macedonia. Only Alaska, Hawaii, Massachusetts, and the District of Columbia were not represented. Those states which represented 10 or more lead authors included: Ohio (29), Pennsylvania (25), Iowa (12), Virginia (12), Nebraska (11), Texas (11), Colorado (11), Oklahoma (10), and California (10). In addition, a maximum of two primary occupations were recorded for each lead author to get a sense of who published the articles. For example, a lead author was frequently listed as both an Extension Specialist and university faculty.

Table 6 was created so that the exact frequency and percentage of articles was reported for each occupation; thus, summing the total frequencies will exceed 302 and summing the total percentages will exceed 100%. Each of the individual occupations or positions listed for the lead author were presented in Table 6 which was organized to show Extension professionals followed by university personnel, other, and unknown. The most frequently listed occupations for the lead author were university faculty (35.1%) and Extension Specialists (24.5%). Those who identified themselves as
employed by Extension (Extension Professionals) made up 47.7% of all lead authors. It is very likely that Extension affiliation would also be high for those lead authors in the ‘unknown’ category.

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extension Agent/Educator</td>
<td>33</td>
<td>10.9</td>
</tr>
<tr>
<td>Extension Specialist</td>
<td>77</td>
<td>24.5</td>
</tr>
<tr>
<td>Extension Director/Leader/Head/Administrator</td>
<td>28</td>
<td>9.3</td>
</tr>
<tr>
<td>Extension Staff</td>
<td>6</td>
<td>2.0</td>
</tr>
<tr>
<td>University Faculty</td>
<td>106</td>
<td>35.1</td>
</tr>
<tr>
<td>University Student</td>
<td>9</td>
<td>3.0</td>
</tr>
<tr>
<td>Other</td>
<td>46</td>
<td>15.2</td>
</tr>
<tr>
<td>Unknown</td>
<td>32</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Table 6: Occupations of Lead Authors for Articles Selected during Stage I

The evaluation object was defined as whatever is being evaluated. The various forms of programming (objects) investigated for this study included the following: (1) individual programs for clientele, (2) educational tools for clientele, (3) individual programs for personnel, (4) the program area, and (5) the U.S./State Cooperative Extension System (CES). The researcher collected data regarding the object of evaluation to determine if there were any differences among forms of programming for the ordinal data collected for the level of evidence documented according to Bennett’s Hierarchy and the UWEX logic model. The Kruskal-Wallis test was selected to calculate mean ranks because it is a nonparametric test to be used when the assumptions of the ANOVA are violated. Mean ranks are a method of handling data which has the same observed frequency occurring at two or more consecutive ranks; it consists of assigning the average of the ranks as the rank for the common frequency.
The mean ranks showed differences among the various objects of the evaluation. The only major practical significance was that the differences for evaluations of the entire U.S./State Cooperative Extension System were lower. As shown in Table 7, the mean rank was 72.27 for Bennett’s Hierarchy and 56.68 for the UWEX logic model.

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Mean Rank for Level-Bennett’s Hierarchy</th>
<th>Mean Rank for Level-UWEX Logic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Programs for Clientele</td>
<td>209</td>
<td>69.2</td>
<td>166.05</td>
<td>166.11</td>
</tr>
<tr>
<td>Educational Tools for Clientele</td>
<td>40</td>
<td>13.3</td>
<td>116.49</td>
<td>116.88</td>
</tr>
<tr>
<td>Individual Programs for Personnel</td>
<td>17</td>
<td>5.6</td>
<td>150.21</td>
<td>150.09</td>
</tr>
<tr>
<td>Program Area</td>
<td>25</td>
<td>8.3</td>
<td>121.62</td>
<td>127.46</td>
</tr>
<tr>
<td>U.S./State CES</td>
<td>11</td>
<td>3.6</td>
<td>72.27</td>
<td>56.68</td>
</tr>
<tr>
<td>Total</td>
<td>302</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Frequencies, Percentages, & Mean Rank for Objects of the Evaluation

Objective 1: To determine the number of program evaluations with outcome data published in the *Journal of Extension* during the years investigated and the program areas represented in these evaluation studies.

There were 302 (24.2%) articles selected, which met the established criteria to be considered evaluative studies with outcome data, out of 1,246 total articles published in the *Journal of Extension* under the headings of ‘Feature Articles’ and ‘Research in Brief’ during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. The frequencies and percentages are displayed in both Table 8 and Figure 9.
<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected</td>
<td>302</td>
<td>24.2</td>
</tr>
<tr>
<td>Not Selected</td>
<td>944</td>
<td>75.8</td>
</tr>
<tr>
<td>Total</td>
<td>1,246</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 8: Articles Reviewed for this Study

Each of the various program areas represented in an article were recorded; thus, individual articles may consist of one or more program areas. Therefore, if the figures listed in Table 9 are added together, the frequencies will be greater than 302 and the percentages will be greater than 100%. The program area of Family and Consumer Sciences was represented the most with 41.1% of all evaluative articles involving this area. Agriculture and Natural Resources followed closely at 39.7% of the articles. The program area of 4-H Youth Development was represented in 31.1% of the articles. Community and Economic Development was represented the least with 15.6% of the articles.
Table 9: Program Areas Represented in Selected Articles

<table>
<thead>
<tr>
<th>Program Area</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family and Consumer Sciences</td>
<td>124</td>
<td>41.1</td>
</tr>
<tr>
<td>Agriculture and Natural Resources</td>
<td>120</td>
<td>39.7</td>
</tr>
<tr>
<td>4-H Youth Development</td>
<td>94</td>
<td>31.1</td>
</tr>
<tr>
<td>Community and Economic Development</td>
<td>47</td>
<td>15.6</td>
</tr>
</tbody>
</table>

Objective 2: To review a subset of articles published in the *Journal of Extension* from 1965-2009 in order to determine the number and percentage of evaluation studies that reached each level of evidence of impact according to both Bennett’s Hierarchy and the logic model.

The frequencies and percentages were displayed in Table 10 for Bennett’s Hierarchy and in Table 11 for the UWEX logic model. The most frequently reported level of evidence measured for Bennett’s Hierarchy and the UWEX logic model was KASA change and short-term outcomes respectively (level 5) for 29.8% of the articles. Bennett’s reaction level data (level 4) corresponds with both level 3 and level 4 of the UWEX logic model; therefore, those percentages reported in Table 11 were lower than those seen in Table 10. The majority of articles were measuring outcomes (levels 5-7) in addition to outputs, with 62.6% of articles reaching these higher levels of impact.
Table 10: Highest Level of Evidence Obtained for Bennett’s Hierarchy

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td>8</td>
<td>2.6</td>
</tr>
<tr>
<td>Activities</td>
<td>10</td>
<td>3.3</td>
</tr>
<tr>
<td>People Involvement</td>
<td>17</td>
<td>5.6</td>
</tr>
<tr>
<td>Reactions</td>
<td>78</td>
<td>25.9</td>
</tr>
<tr>
<td>KASA Change</td>
<td>90</td>
<td>29.8</td>
</tr>
<tr>
<td>Practice Change</td>
<td>82</td>
<td>27.2</td>
</tr>
<tr>
<td>End Results</td>
<td>17</td>
<td>5.6</td>
</tr>
<tr>
<td>Total</td>
<td>302</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 11: Highest Level of Evidence Obtained for UWEX Logic Model

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation/Priorities</td>
<td>2</td>
<td>0.7</td>
</tr>
<tr>
<td>Inputs</td>
<td>6</td>
<td>2.0</td>
</tr>
<tr>
<td>Outputs-Activities</td>
<td>55</td>
<td>18.2</td>
</tr>
<tr>
<td>Outputs-Participation</td>
<td>50</td>
<td>16.6</td>
</tr>
<tr>
<td>Short-Term Outcomes</td>
<td>90</td>
<td>29.8</td>
</tr>
<tr>
<td>Medium-Term Outcomes</td>
<td>82</td>
<td>27.2</td>
</tr>
<tr>
<td>Long-Term Outcomes</td>
<td>17</td>
<td>5.6</td>
</tr>
<tr>
<td>Total</td>
<td>302</td>
<td>100.1</td>
</tr>
</tbody>
</table>

Table 12 displayed the percentiles according to the level of evidence. Twenty-five percent of the articles documented evidence of at least practice change; however, 25% also documented evidence of impact below Bennett’s level of reaction. Please note in Table 12 that levels 5, 6, and 7 are the same between models, but there are differences at the lower levels. For example, level 4 of Bennett’s Hierarchy corresponds to either levels 3 or 4 of the UWEX logic model depending on the type of reaction data. The rest of Bennett’s levels correspond to the UWEX logic model as follows: level 3 (participation) ~ level 4, level 2 (activities) ~ level 3, and level 1 (inputs) ~ level 2.
<table>
<thead>
<tr>
<th>Level-Bennett’s Hierarchy</th>
<th>Percentiles</th>
<th>Level-UWEX logic model</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>50&lt;sup&gt;th&lt;/sup&gt; (median)</td>
<td>75&lt;sup&gt;th&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>4 (reactions)</td>
<td>5 (KASA change)</td>
<td>6 (practice change)</td>
</tr>
<tr>
<td></td>
<td>4 (outputs-participation)</td>
<td>5 (short-term outcomes)</td>
<td>6 (medium-term outcomes)</td>
</tr>
</tbody>
</table>

Table 12: Percentile of articles According to Level of Evidence

**Objective 3: To investigate and identify any trends over time.**

As reported in Tables 13 and 14, the number of articles meeting the established criteria to be considered as evaluative studies with outcome data has increased over the years going from 16 articles in the 1960s to 150 articles in the 2000s. There were a higher percentage of evaluative studies stopping at the lower levels of evidence in the 1960s and 1970s. However, for KASA change and short-term outcomes, percentages have remained fairly consistent over time with a dip in the 1980s. Practice change and medium-term outcomes were most frequently reported in the 2000s. There has been much less documentation of end results and long-term outcomes throughout the years.

Figures 10 and 11 provide a different way to visualize the changes throughout the decades. Ideally, the figures should display an incline both over the decades and over the levels of evidence. There was a very noticeable incline over the years; however, it was easy to see that KASA change/short-term outcomes were most predominant in each of the decades with the exception of the 1980s. The levels were more bell-shaped instead of incline-shaped. However, in the recent decades, the bell-shape was more negatively skewed which represents higher evidence of impact.
Table 13: Bennett’s Hierarchy by Decade (frequency [percent])

<table>
<thead>
<tr>
<th>Decade</th>
<th>Inputs</th>
<th>Activities</th>
<th>People Involvement</th>
<th>Reactions</th>
<th>KASA Change</th>
<th>Practice Change</th>
<th>End Results</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>2 (12.5)</td>
<td>2 (12.5)</td>
<td>3 (18.8)</td>
<td>3 (18.8)</td>
<td>5 (31.3)</td>
<td>0 (0.0)</td>
<td>1 (6.3)</td>
<td>16</td>
</tr>
<tr>
<td>1970s</td>
<td>1 (6.3)</td>
<td>3 (18.8)</td>
<td>0 (0.0)</td>
<td>3 (18.8)</td>
<td>5 (31.3)</td>
<td>4 (25.5)</td>
<td>0 (0.0)</td>
<td>16</td>
</tr>
<tr>
<td>1980s</td>
<td>2 (4.9)</td>
<td>2 (4.9)</td>
<td>2 (4.9)</td>
<td>14 (34.2)</td>
<td>5 (12.2)</td>
<td>14 (12.2)</td>
<td>2 (4.9)</td>
<td>41</td>
</tr>
<tr>
<td>1990s</td>
<td>1 (1.3)</td>
<td>3 (3.8)</td>
<td>2 (2.5)</td>
<td>24 (30.4)</td>
<td>27 (34.2)</td>
<td>18 (22.8)</td>
<td>4 (5.1)</td>
<td>79</td>
</tr>
<tr>
<td>2000s</td>
<td>2 (1.3)</td>
<td>0 (0.0)</td>
<td>10 (6.7)</td>
<td>34 (22.7)</td>
<td>48 (30.0)</td>
<td>46 (30.0)</td>
<td>10 (6.7)</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>8 (2.6)</td>
<td>10 (3.3)</td>
<td>17 (5.6)</td>
<td>78 (25.9)</td>
<td>90 (29.8)</td>
<td>82 (27.2)</td>
<td>17 (5.6)</td>
<td>302</td>
</tr>
</tbody>
</table>

Figure 10: Bennett’s Hierarchy by Decade
<table>
<thead>
<tr>
<th>Decade</th>
<th>Situation/ Priorities</th>
<th>Inputs</th>
<th>Outputs- Activities</th>
<th>Outputs- Participation</th>
<th>Short-Term Outcomes</th>
<th>Medium-Term Outcomes</th>
<th>Long-Term Outcomes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>1 (6.3)</td>
<td>1 (6.3)</td>
<td>2 (12.5)</td>
<td>5 (31.3)</td>
<td>5 (31.3)</td>
<td>0 (0.0)</td>
<td>1 (6.3)</td>
<td>16</td>
</tr>
<tr>
<td>1970s</td>
<td>0 (0.0)</td>
<td>1 (6.3)</td>
<td>5 (31.3)</td>
<td>1 (6.3)</td>
<td>5 (31.3)</td>
<td>4 (25.5)</td>
<td>0 (0.0)</td>
<td>16</td>
</tr>
<tr>
<td>1980s</td>
<td>0 (0.0)</td>
<td>2 (4.9)</td>
<td>10 (24.4)</td>
<td>8 (19.5)</td>
<td>5 (12.2)</td>
<td>14 (12.2)</td>
<td>2 (4.9)</td>
<td>41</td>
</tr>
<tr>
<td>1990s</td>
<td>0 (0.0)</td>
<td>1 (1.3)</td>
<td>16 (20.3)</td>
<td>13 (16.5)</td>
<td>27 (34.2)</td>
<td>18 (22.8)</td>
<td>4 (5.1)</td>
<td>79</td>
</tr>
<tr>
<td>2000s</td>
<td>1 (0.7)</td>
<td>1 (0.7)</td>
<td>22 (14.7)</td>
<td>22 (14.7)</td>
<td>48 (32.0)</td>
<td>46 (30.0)</td>
<td>10 (6.7)</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>2 (0.7)</td>
<td>6 (2.0)</td>
<td>55 (18.2)</td>
<td>50 (16.6)</td>
<td>90 (29.8)</td>
<td>82 (27.2)</td>
<td>17 (5.6)</td>
<td>302</td>
</tr>
</tbody>
</table>

Table 14: UWEX Logic Model by Decade (frequency [percent])

Figure 11: UWEX Logic Model by Decade
Another major observable difference was that a larger percentage of articles stopped collecting evidence at the lower levels of Bennett’s Hierarchy (inputs, activities, & people involvement) and the lower levels of the UWEX Logic Model during the 1960s. This helps to show that a much greater percentage of articles were documenting higher-level evidence of impact in the more recent decades. Tables 15 and 16 used groupings to more clearly show that a greater frequency and percentage of articles were documenting higher-value evidence (outcomes) over time.

<table>
<thead>
<tr>
<th>Decade</th>
<th>Inputs</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>2 (12.5)</td>
<td>8 (50)</td>
<td>6 (37.5)</td>
<td>16</td>
</tr>
<tr>
<td>1970s</td>
<td>1 (6.3)</td>
<td>6 (37.5)</td>
<td>9 (56.3)</td>
<td>16</td>
</tr>
<tr>
<td>1980s</td>
<td>2 (4.9)</td>
<td>18 (43.9)</td>
<td>21 (51.2)</td>
<td>41</td>
</tr>
<tr>
<td>1990s</td>
<td>1 (1.3)</td>
<td>29 (36.7)</td>
<td>49 (62.0)</td>
<td>79</td>
</tr>
<tr>
<td>2000s</td>
<td>2 (1.3)</td>
<td>44 (29.3)</td>
<td>104 (69.3)</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>8 (2.6)</td>
<td>105 (34.8)</td>
<td>189 (62.6)</td>
<td>302</td>
</tr>
</tbody>
</table>

Table 15: Bennett’s Hierarchy Groupings by Decade (frequency [percent])

<table>
<thead>
<tr>
<th>Decade</th>
<th>Situation/ Priorities</th>
<th>Inputs</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>1 (6.3)</td>
<td>1 (6.3)</td>
<td>7 (43.8)</td>
<td>6 (37.5)</td>
<td>16</td>
</tr>
<tr>
<td>1970s</td>
<td>0 (0.0)</td>
<td>1 (6.3)</td>
<td>6 (37.5)</td>
<td>9 (56.3)</td>
<td>16</td>
</tr>
<tr>
<td>1980s</td>
<td>0 (0.0)</td>
<td>2 (4.9)</td>
<td>18 (43.9)</td>
<td>21 (51.2)</td>
<td>41</td>
</tr>
<tr>
<td>1990s</td>
<td>0 (0.0)</td>
<td>1 (1.3)</td>
<td>29 (36.7)</td>
<td>49 (62.0)</td>
<td>79</td>
</tr>
<tr>
<td>2000s</td>
<td>1 (0.7)</td>
<td>1 (0.7)</td>
<td>44 (29.3)</td>
<td>104 (69.3)</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>2 (0.7)</td>
<td>6 (2.0)</td>
<td>105 (34.8)</td>
<td>189 (62.6)</td>
<td>302</td>
</tr>
</tbody>
</table>

Table 16: UWEX Logic Model Groupings by Decade (frequency [percent])

For substantial impact to occur, Diem (1997; 2003) stated that the results of a program must ultimately change people’s attitudes or behavior, or benefit society in other ways. Changes in attitude were documented at level 5 (KASA change/short-term outcomes), but a change in attitude does not necessarily show application. Therefore,
more substantial evidence of impact begins at level 6 (practice change/medium-term outcomes) because that is where the results of programming are actually applied. Table 17 displays the frequency and percentage of articles reporting more substantial evidence of impact which included both level 6 (practice change/medium-term outcomes) and level 7 (end results/long-term outcomes).

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>1</td>
<td>6.3</td>
</tr>
<tr>
<td>1970s</td>
<td>4</td>
<td>25.0</td>
</tr>
<tr>
<td>1980s</td>
<td>16</td>
<td>39.0</td>
</tr>
<tr>
<td>1990s</td>
<td>22</td>
<td>27.9</td>
</tr>
<tr>
<td>2000s</td>
<td>56</td>
<td>37.3</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>32.8</td>
</tr>
</tbody>
</table>

Table 17: Frequency & Percentage of Articles Documenting Evidence of Substantial Impact According to Diem (2003)

The Kruskal-Wallis test was used with ordinal data to determine if there were differences between the decades. Table 18 displays the calculated mean ranks as well as the frequencies and percentages of articles. Figure 12 allows one to see the trend over the years as well as the leveling that occurred during the 1980s. In addition, Figure 12 shows more separation between the models during the early decades as lower levels of evidence were more frequently documented and the models differ much more at those lower levels.
Objective 4: To identify and quantify the primary evaluation and data collection methods used to document evidence of impact.

The evaluation methods and data collection methods were taken and slightly modified from Diem’s (2003) examples of basic methods and tools used to measure impact. A maximum of two primary evaluation methods and two primary data collection methods were recorded on the instrument; thus, there was one or more methods listed for
each article. The frequencies and percentages for each method were reported in Tables 19 and 20.

The primary evaluation methods utilized were reported in Table 19. Most of the articles reviewed for this study (74.8%) used an evaluation method which examined the program by asking others (survey).

<table>
<thead>
<tr>
<th>Method</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask them</td>
<td>226</td>
<td>74.8</td>
</tr>
<tr>
<td>Test them</td>
<td>91</td>
<td>30.1</td>
</tr>
<tr>
<td>Observe them</td>
<td>33</td>
<td>10.9</td>
</tr>
</tbody>
</table>

Table 19: Primary Evaluation Methods Utilized

The primary data collection methods were reported in Table 20. Each of the methods listed on the instrument were observed in this study. There were 50.3% of the articles that utilized some form of a written questionnaire. The next most frequent methods used were pre-tests/post-tests (20.2%) and interviews/testimonials/case studies/focus groups (17.9%). The remainder of the data collection methods were utilized in less than 9% of the articles.

<table>
<thead>
<tr>
<th>Method</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written questionnaires</td>
<td>152</td>
<td>50.3</td>
</tr>
<tr>
<td>Pre-test/post-test</td>
<td>61</td>
<td>20.2</td>
</tr>
<tr>
<td>Interviews/testimonials/case studies/focus groups</td>
<td>54</td>
<td>17.9</td>
</tr>
<tr>
<td>Post-test with control group comparison</td>
<td>25</td>
<td>8.3</td>
</tr>
<tr>
<td>Retrospective pretest / post-then-pre</td>
<td>16</td>
<td>5.3</td>
</tr>
<tr>
<td>Direct observation</td>
<td>15</td>
<td>5.0</td>
</tr>
<tr>
<td>Reviewing information from other sources</td>
<td>14</td>
<td>4.6</td>
</tr>
<tr>
<td>Online survey</td>
<td>14</td>
<td>4.6</td>
</tr>
<tr>
<td>Longitudinal studies</td>
<td>3</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 20: Primary Data Collection Methods Utilized
Table 21 displays the frequency that each data collection method was used to document evidence for each of Bennett’s levels. In addition, it shows the percentage by column that each method was used for a specific level of evidence of impact according to Bennett’s levels. The written questionnaire was the most frequently used data collection method for inputs (level 1), people involvement (level 3), reactions (level 4), practice change (level 6), and end results (level 7). Reviewing information from other sources was the most frequently used for activities (level 2), and pre-test/post-test was the most frequently used for KASA change (level 5).

<table>
<thead>
<tr>
<th>Method</th>
<th>Inputs (level 1)</th>
<th>Activities (level 2)</th>
<th>People Invol. (level 3)</th>
<th>Reactions (level 4)</th>
<th>KASA Change (level 5)</th>
<th>Practice Change (level 6)</th>
<th>End Results (level 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written questionnaires</td>
<td>4 (50%)</td>
<td>2 (20%)</td>
<td>6 (31.6%)</td>
<td>59 (67.1%)</td>
<td>30 (28.0%)</td>
<td>43 (42.2%)</td>
<td>8 (40%)</td>
</tr>
<tr>
<td>Longitudinal study</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (1.1%)</td>
<td>1 (0.9%)</td>
<td>1 (0.9%)</td>
<td>0</td>
</tr>
<tr>
<td>Interviews/testimonials/case studies/focus groups</td>
<td>1 (12.5%)</td>
<td>2 (20%)</td>
<td>2 (10.5%)</td>
<td>14 (15.9%)</td>
<td>13 (12.2%)</td>
<td>18 (17.7%)</td>
<td>4 (20%)</td>
</tr>
<tr>
<td>Pre-test/post-test</td>
<td>1 (12.5%)</td>
<td>1 (10%)</td>
<td>2 (10.5%)</td>
<td>2 (2.3%)</td>
<td>32 (30.0%)</td>
<td>23 (22.6%)</td>
<td>0</td>
</tr>
<tr>
<td>Post-test with control group comparison</td>
<td>1 (12.5%)</td>
<td>0</td>
<td>0</td>
<td>2 (2.3%)</td>
<td>14 (13.1%)</td>
<td>7 (6.9%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Direct observation</td>
<td>1 (12.5%)</td>
<td>2 (20%)</td>
<td>3 (15.8%)</td>
<td>4 (4.6%)</td>
<td>3 (2.8%)</td>
<td>1 (1.0%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Reviewing information from other sources</td>
<td>0</td>
<td>3 (30%)</td>
<td>4 (21.1%)</td>
<td>0</td>
<td>1 (0.9%)</td>
<td>2 (2.0%)</td>
<td>4 (20%)</td>
</tr>
<tr>
<td>Online survey</td>
<td>0</td>
<td>0</td>
<td>2 (10.5%)</td>
<td>6 (6.8%)</td>
<td>3 (2.8%)</td>
<td>2 (2.0%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Retrospective pretest/post-then-pre</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10 (9.4%)</td>
<td>5 (4.9%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8 (100%)</strong></td>
<td><strong>10 (100%)</strong></td>
<td><strong>19 (100%)</strong></td>
<td><strong>88 (100%)</strong></td>
<td><strong>107 (100%)</strong></td>
<td><strong>102 (100%)</strong></td>
<td><strong>20 (100%)</strong></td>
</tr>
</tbody>
</table>

Table 21: Frequency and percentage that each data collection method was used for each of Bennett’s levels.
Objective 5: To determine if the program being evaluated was designed for reach (number of people interacted with) and/or richness (detail and depth of interaction).

Reach was defined as the amount of individuals you can touch and influence through programming (King & Boehlje, 2000). Richness was defined as the depth and detail of your interaction (King & Boehlje, 2000). Both concepts were operationally defined as either high or low. There was no medium or range examined for the exploratory purposes of this study; thus, the researcher was required to select either high or low for each concept.

As shown in Table 22, there were 157 (52.0%) articles coded as low reach and 145 (48.0%) coded as high reach. In addition, there were 85 (28.2%) articles coded as low richness and 217 (71.9%) coded as high richness. Table 22 also displayed the frequency of the four possible combinations of reach and richness. There were three combinations of reach and richness that were observed for the 302 articles investigated during Stage II (reading the articles and coding the data). No articles were coded as low reach/low richness. The combination of low reach/high richness consisted of 52.0% of the articles, high reach/low richness consisted of 28.1% of the articles, and high reach/high richness consisted of 19.9% of the articles.

<table>
<thead>
<tr>
<th>Level of Reach</th>
<th>Level of Richness</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>157 (52.0)</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>60 (19.9)</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>85 (28.2)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 22: Cross Tabulation of Reach and Richness (frequency [percent])
Objective 6: To determine the geographic scope (county [local], multicounty regional, statewide, multistate regional, or national) of the programs evaluated.

The geographic scope was defined as the region/area or location of the clientele/audience targeted for the programming. It was important to note that the geographic scope had no bearing on the level of reach, as reach pertains to the delivery of information. Each of the variables was defined differently and any program of any geographic scope may have either high or low reach.

Geographic scope was operationally defined for this study by using five ordinal categories. The categories were as follows: county (local), multicounty regional, statewide, multistate regional, or national based on the program represented in the article/study. The frequency and percentage of articles for each level of geographic scope was presented in Table 23. Most of the programs represented in this study were statewide in scope (51.0%). The remaining geographic scopes were represented in less than 20% of the articles. Multistate and national programs were represented in less than 7% of the articles.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>County (local)</td>
<td>56</td>
</tr>
<tr>
<td>Multicounty regional</td>
<td>58</td>
</tr>
<tr>
<td>Statewide</td>
<td>154</td>
</tr>
<tr>
<td>Multistate regional</td>
<td>19</td>
</tr>
<tr>
<td>National</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>302</td>
</tr>
</tbody>
</table>

Table 23: Geographic Scope of Programs Evaluated
Objective 7: To explore the relationship between reach/richness, geographic scope, and the level of evaluation according to both Bennett’s Hierarchy and the logic model.

The nonparametric Spearman rank-order correlation coefficient was utilized to explore the magnitude and direction of the associations. The magnitude of the relationship was interpreted using Bartz’s (1999) terminology as follows: very high (0.80 or higher), strong (0.60 to 0.80), moderate (0.40 to 0.60), low (0.20 to 0.40), and very low (0.20 or lower).

Each of the ordinal variables shown in Table 24 were coded as follows: Level-Bennett’s Hierarchy ranged from 1 to 7 (i.e., 1 = inputs, 2 = activities, 3 = people involvement, 4 = reactions, 5 = KASA change, 6 = practice change, 7 = end results); Level-University of Wisconsin-Extension logic model format also ranged from 1 to 7 (i.e., 1 = situation/priorities, 2 = inputs, 3 = outputs-activities, 4 = outputs-participation, 5 = short-term outcomes, 6 = medium-term outcomes, 7 = long-term outcomes); both reach and richness were recorded separately as either 1 (low) or 2 (high); and geographic scope ranged from 1 to 5 (i.e., 1 = county [local], 2 = multicounty regional, 3 = statewide, 4 = multistate regional, and 5 = national).

The correlations were displayed in Table 24. As expected, the level of Bennett’s Hierarchy was very highly associated with the level of the UWEX logic model as the three highest levels (levels 5-7) perfectly correspond with each other. The only difference between the two variables occurred at the lower levels (level 1-4). Also as expected, there was a strong negative relationship with reach and richness. The association between Bennett’s Hierarchy/UWEX logic model and the level of reach was a
low negative relationship. The association between Bennett’s Hierarchy/UWEX logic model and the level of richness was a very low positive relationship. The association between Bennett’s Hierarchy/UWEX logic model and geographic was also a very low positive relationship. In addition, there were very low relationships between geographic scope and both level of reach and level of richness.

<table>
<thead>
<tr>
<th></th>
<th>Level-Bennett's Hierarchy</th>
<th>Level-UWEX Logic Model</th>
<th>Level of Reach</th>
<th>Level of Richness</th>
<th>Geographic Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-Bennett's Hierarchy</td>
<td>1.000</td>
<td>0.967</td>
<td>-0.275</td>
<td>0.124</td>
<td>-0.030</td>
</tr>
<tr>
<td>Level-UWEX Logic Model</td>
<td>0.967</td>
<td>1.000</td>
<td>-0.293</td>
<td>0.131</td>
<td>-0.067</td>
</tr>
<tr>
<td>Level of Reach</td>
<td>-0.275</td>
<td>-0.293</td>
<td>1.000</td>
<td>-0.651</td>
<td>0.142</td>
</tr>
<tr>
<td>Level of Richness</td>
<td>0.124</td>
<td>0.131</td>
<td>-0.651</td>
<td>1.000</td>
<td>0.037</td>
</tr>
<tr>
<td>Geographic Scope</td>
<td>-0.030</td>
<td>-0.067</td>
<td>0.142</td>
<td>0.037</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 24: Correlation Table showing Spearman’s rho
CHAPTER 5
SUMMARY, DISCUSSION, & CONCLUSIONS

Summary

The purpose of this study was to explore and describe the level of evidence of impact collected through program evaluation (outcome studies) by the Cooperative Extension System as published in the *Journal of Extension*; and to examine any potential relationships between the documentation of evidence of impact with reach, richness, and geographic scope. Impact was operationally defined as program effectiveness determined by the highest level of evidence collected based on the components of both Bennett’s Hierarchy (i.e., 1 = inputs, 2 = activities, 3 = people involvement, 4 = reactions, 5 = KASA change, 6 = practice change, 7 = end results) and the UWEX logic model format (i.e., 1 = situation/priorities, 2 = inputs, 3 = outputs-activities, 4 = outputs-participation, 5 = short-term outcomes, 6 = medium-term outcomes, 7 = long-term outcomes). Articles reviewed in the *Journal of Extension* were those listed under the headings of ‘Feature Articles’ and ‘Research in Brief’ during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. To achieve this purpose, a set of research objectives was created to guide the investigation.

This study utilized a form of quantitative content analysis where data was collected and coded using qualitative methods and analyzed via quantitative techniques.
The articles were reviewed qualitatively and the instrument was completed by making a professional judgment and coding the appropriate choice for each item based on predetermined criteria contained in a rubric. Quantitative data analysis procedures were used to interpret the coded data.

The instrumentation procedures utilized a researcher-completed observation form or tally sheet (Fraenkel & Wallen, 2006). The instrument consisted of a Microsoft® Office Word® 2007 document (see Appendix B) to use as a questionnaire and a Microsoft® Office Excel® 2007 spreadsheet (see Appendix C) to use for entering data. In addition, the rubric (see Appendix A) helped to ensure that data was collected accurately and consistently.

The primary variables used for the analyses included the following levels: Level-Bennett’s Hierarchy ranged from 1 to 7 (e.g., 1 = inputs, 2 = activities, 3 = people involvement, 4 = reactions, 5 = KASA change, 6 = practice change, 7 = end results); Level-University of Wisconsin-Extension logic model format also ranged from 1 to 7 (e.g., 1 = situation/priorities, 2 = inputs, 3 = outputs-activities, 4 = outputs-participation, 5 = short-term outcomes, 6 = medium-term outcomes, 7 = long-term outcomes); both reach and richness were recorded separately as either 1 (low) or 2 (high); and geographic scope ranged from 1 to 5 (e.g., 1 = county [local], 2 = multicounty regional, 3 = statewide, 4 = multistate regional, and 5 = national).

The formal data collection consisted of a two-stage process. Stage I (selection of articles) involved screening abstracts and scanning articles to select those that were appropriate based on the criteria established for this study. Stage II (collection of data)
involved reading the selected articles, coding the text, and completing the remainder of the instrument.

**Summary of Limitations.** There were some limitations that were recognized for this study. First, many program evaluations conducted for Extension programs were likely published and printed in various other journals and publications. The *Journal of Extension* was selected for the purposes of this study because it was reported to be the official refereed journal of the U.S. Cooperative Extension System (Extension Journal, Inc., 2010). Another aspect related to this limitation involved examining only those outcome evaluation results that were actually published. It is likely that many programs were formally evaluated with findings that were utilized for accountability and/or improvement, but the results were never formally published.

A second limitation resulted from the fact that only articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’ were reviewed for this study. Additional articles reporting on outcome studies and the documentation of evidence of impact were likely included under other headings, specifically ‘Ideas at Work’ and ‘Tools of the Trade’; but these articles were typically shorter and less comprehensive, often missing pertinent information required for the purposes of this study. Another aspect related to this limitation was that the researcher was also limited to the actual information contained within the individual articles published under the headings of ‘Feature Articles’ and ‘Research in Brief’.

A third limitation involved the *Journal of Extension* itself. The journal has evolved and changed policies and procedures throughout its existence. There were likely
different requirements for publishing articles in the 1960s than in the 2000s. A similar limitation involved the PDF of the article as presented on the JOE web site. Sometime the scanning quality was poor and a portion of the article appeared very dark or distorted. This required more careful attention and educated interpretations in reading the distorted words.

Finally, an additional limitation involved the clientele who published information in the Journal of Extension. Agent/Educators and other Extension personnel may or may not have a tenure-track appointment with publication requirements. Those without publication requirements might be less likely to publish evaluation results. Those specialists located on campus with tenure-track appointments may be more apt to publish in other research journals more specific to their field of expertise. State Extension Specialists are often responsible for major statewide programs which could be more likely to be formally evaluated with published results than smaller, localized programs.

Summary and Discussion of Findings

The findings of this study were organized and reported by describing the articles selected and addressing each of the research objectives. There were a total of 1,246 articles investigated in which 302 were selected during Stage I. The relevant articles selected during Stage I were written by lead authors who represented 47 of the 50 states. Based on this, one can infer that these evaluative studies represented the entire U.S. Cooperative Extension System. The lead authors were predominantly Extension personnel if it is assumed that a portion of those not reporting an occupation were also
Extension personnel. Roughly one-third of the lead authors considered themselves as university faculty. Extension personnel may or may not be considered as university faculty based on the criteria and structure used in a particular state’s system.

**Objective 1: To determine the number of program evaluations with outcome data published in the Journal of Extension during the years investigated and the program areas represented in these evaluation studies.**

**Summary.** There were 302 articles selected, which met the established criteria to be considered evaluative studies with outcome data, out of 1,246 total articles published in the *Journal of Extension* under the headings of ‘Feature Articles’ and ‘Research in Brief’ during 1965-69, 1975-79, 1985-89, 1995-99, and 2005-09. There were 24.2% of the articles which were evaluative studies with outcome data. Family and Consumer Science was represented most frequently followed very closely by Agriculture and Natural Resources. The program area of 4-H Youth Development was also well represented with almost a third of the articles. Community and Economic Development was represented the least.

**Discussion.** Based on the findings of objective 1, it is likely that around one-quarter of all published articles fit the criteria of evaluative outcome studies; but this cannot be inferred as the articles reviewed were not randomly selected. Furthermore, only the *Journal of Extension* was used. Evaluative articles were likely printed and published in other journals and outlets. The purpose of the *Journal of Extension* is much broader than reporting results from program evaluation; therefore, the fact that one-fourth
of the articles were reporting evaluative results can be interpreted both positively and negatively. Some individuals might believe that more evaluative articles are needed while others might be very satisfied that three-fourths of the articles pertain to needs assessment or information with implications for Extension programming.

A lack of evaluation skills and/or lower priorities might explain this finding. Extension personnel have never been overly interested in developing evaluation skills (Archer, Bruns, & Heaney, 2007). Archer, Bruns, and Heaney (2007) identified two potential reasons for this. First, most Extension employees are hired because of their discipline expertise and their interest in teaching. Second, evaluation is not a part of their academic arsenal, and it may not be as high of a priority when it comes to professional development. With limited resources involving personnel and money, Extension professionals tend to do more programming and less evaluation (Archer, Bruns, & Heaney, 2007).

Family and Consumer Sciences was found to be the most frequently represented program area. However, Agriculture and Natural Resources and 4-H Youth Development were still very well represented. Community and Economic Development was represented the least. This program area was also the youngest of the major program areas and likely the least represented in the organization.

Traditionally, personnel in a specific program area tended to work within his or her area of responsibility; thus, it would be expected that each program area would be represented in around 25% to 35% of the articles just by chance. Today, personnel tend to collaborate and work across program areas in cross-disciplinary teams to address
specific problems or issues. For example, many problems and issues pertain to both adults and youth; thus, the 4-H Youth Development expert is often involved with the youth component of that programming. In addition, Agents/Educators have become more integrated and focused on specific clientele groups. Clientele can be categorized as adults, youth, families, and communities. Each of the four program areas often involves subject matter that was relevant to each of these groups. It is likely that multiple program areas will be more represented in future programming. Although, as individuals take on multiple program responsibilities (often due to budget cuts), the pattern may change. Objective 3 provides the trends of these results for examining the findings over time.

Objective 2: To review a subset of articles published in the Journal of Extension from 1965-2009 in order to determine the number and percentage of evaluation studies that reached each level of evidence of impact according to both Bennett’s Hierarchy and the logic model.

Summary. The most frequent level of evaluation reported was level 5 (KASA change/short-term outcomes) at 29.8%. Level 6 (practice change) was the next most frequent at 27.2%. Levels 5, 6, and 7 represent outcomes. Those levels below level 5 represent either outputs, inputs, or situation/priorities. The findings indicated that a majority (almost two-thirds) were at least measuring outcomes. Levels 6 and 7 represent more substantial evidence of impact. The findings indicated that about one-third of the articles were measuring substantial evidence of impact.
**Discussion.** Too often, Extension programs do not document impact by collecting real evidence of behavior change and/or greater end results that benefit society through social, economic, civic, and environmental improvement. Many Extension personnel consider programs successful in lower levels of the program progression (Richardson, 2005). Gentry-Van Laanen and Nies (1995) stated that Extension evaluators have primarily reported on the number and demographics of the people who participated in programs. Stup (2003) reported that most program evaluations stop at the reaction level (level 4) or learning (level 5) without measuring higher level changes. Furthermore, Duttweiler (2008) found that 35% of articles from 1998 through 2007 provided some evidence of effectiveness, but it was still unknown as to the percentage of articles that documented evidence of impact.

The findings of this study have provided some answers to the problem; however, a judgment must be made as to whether enough is being accomplished or more is needed. The data collected in this study have shown that 27.2% of the articles regarding evaluative outcome studies were documenting some evidence of behavior change. An additional 5.6% were showing some evidence of greater end results. It is likely that most would agree that substantial impact occurs at the level of behavior change because this is where programming results are actually applied. Prior to this level, tremendous learning might take place, but it only truly makes a significant difference to individuals and society if that learning is applied. A "one-shot" lesson has limited impact if it does not result in some "doing" by the participants; only then does true "knowing" take place (Clements, 1999). Currently, about one-third of published outcome studies were
documenting higher-level or substantial evidence of impact. The point was raised that most outcome studies stop at the level of reaction and learning. Currently, the highest level of evidence obtained was either reaction or KASA change for 55.7% of the articles. Therefore, around two-thirds of the outcome studies were not documenting evidence of practice change or the actual application of knowledge, attitudes, skills, and aspirations.

Another point was made that Extension evaluators were simply reporting the number and demographics of participants. The findings of this study do not support this claim. This investigation has shown that 88.5% of the articles documented evidence above the level of people involvement (level 3) according to Bennett’s Hierarchy. In addition, 62.6% of the articles were above the level of outputs-participation (level 4) according to the UWEX logic model. However, the UWEX logic model is somewhat different than Bennett’s Hierarchy at the lower levels as Bennett’s reaction level data corresponds to either outputs-activities or outputs-participation. Therefore, there is a portion of studies contained in levels 3 and 4 of the UWEX logic model that would be considered as reaching Bennett’s level 4 (reactions).

There is a great deal of subjectivity in the reaction and KASA change levels of Bennett’s Hierarchy as well as the outputs and short-term outcomes of the UWEX logic model format. The United Way (1996) reported that outputs are about the program, while outcomes are about the participants. In reality, reaction level data can include outputs and/or outcomes. Leech, Sutherland, and Wainwright (2004) would consider evidence at the reaction level as an outcome and place it in the short-term outcome section of the logic model. Based on the various publications written by either Claude
Bennett or the University of Wisconsin Extension, reaction level data was primarily considered as an output for purposes of this study. This is because someone’s subjective reaction to programming does not truly provide evidence of KASA change. A pre-test/post-test can provide much greater evidence of KASA change. However, this is not a clear-cut issue. Aspirations and attitudes which can be measured or tested from reactionary data are part of KASA. Jayaratne (2010) described an aspiration as the heightened level of internal motivation of an individual for taking charge of a behavior or practice with full comprehension of its content, value, and application for achieving desired benefits.

As outlined earlier in the significance of the study portion of Chapter 1, it is important for Extension personnel to know what percentage of evaluative studies are providing accountability and performance. If one desires behavior change as the minimum to document evidence of substantial impact, there were one-third of the evaluative articles that either met or exceeded this level. According to Clements (1999), legislators agree that impact means behavior change. Second, to show evidence of accountability, decision makers need to document the outcomes. Outputs are important to program administrators for both formative and summative purposes; however, funders and legislators will be primarily interested in outcomes that make an impact. Based on the findings of this study, evaluative studies were documenting evidence of outcomes about two-thirds of the time. There are two issues associated with this number. Do these studies provide enough evidence for justification, and do these studies make enough significant impact for society? These are judgment questions that will be decided by
administrators and legislators. However, Extension personnel must be able to answer these questions in the affirmative to increase the likelihood that legislators will agree.

**Objective 3: To investigate and identify any trends over time.**

**Summary.** Objective 3 provided similar information as Objective 2, but allowed one to view the findings by decade for trend analysis. This investigation clearly showed that the number of evaluative studies with outcome data has increased over the years going from 16 articles in the 1960s to 150 articles in the 2000s. In more recent decades, the percentage of evaluative studies reaching a specific level of evidence has remained somewhat similar.

**Discussion.** It is important to consider the evolution of both Cooperative Extension and program evaluation. The original Extension Service was formally established by the Smith Lever Act in 1914 (Rasmussen, 1989). The professional journal was not created until 1963 (Extension Journal, Inc., 2010). Extension only had one-half of a century to develop and mature prior to the creation of a professional journal. Those journal articles published in the 1960s were written at a time when the *Journal of Extension* was very new and Cooperative Extension was still relatively young. The *Journal of Extension* has changed and evolved over the years. In the early years, most articles could be described as professional editorials by Extension experts. It was rare to find an article that reported on specific results and findings from the collection of data. There were many abstracts of research studies from other journals or unpublished theses
and dissertations included with implications for Cooperative Extension written by the editor.

As Cooperative Extension has matured, JOE has evolved into a journal with articles and studies that are very similar to those found in other modern-era research journals. Bennett’s Hierarchy was first published in the *Journal of Extension* in 1975. Prior to this, there were some articles and commentaries regarding the need and importance of evaluation in Extension. In September of 1983, an entire JOE issue (volume 21, number 5) was devoted to the practice of program evaluation. The number of evaluative outcome studies resulting from programmatic data has significantly increased over the decades. Today, Bennett’s Hierarchy, as well as other programming models, has essentially been replaced with the logic model. However, based on this study, one should see and understand the importance of using Bennett’s original terminology to help describe evidence of impact when using the logic model.

In addition, the actual practice of program evaluation didn’t formally exist until the 1960s when the U.S. government began allocating substantial funding and resources for social programs (Scriven, 2003). The practice and/or profession of program evaluation was born because the federal government wanted to determine the effectiveness of social initiatives, improve accountability to the American people, and legitimize newly created public programs (Carman, Fredericks, & Introcaso, 2008). Over the past 40 years, program evaluation has become common practice because of the increased demands of accountability. The methods used for program evaluation have been borrowed and adapted from the discipline of social science research. Michael
Scriven (2003) has stated that the profession of program evaluation is moving towards acceptance as a discipline, but still has a long way to go. Because of the value-free doctrine in the standard social sciences, professionals did not view program evaluation as objective and scientific, but this viewpoint changed in the late twentieth century (Scriven, 2003).

A major observable change found from this study was that a larger percentage of articles published during the 1960s simply reported on inputs, activities, and people involvement. In the early years of JOE, it was more uncommon to report outcomes and document higher-level evidence of impact. A larger percentage of articles stopped collecting evidence at the lower levels of Bennett’s Hierarchy (inputs, activities, & people involvement) and the lower levels of the UWEX Logic Model during the 1960s. This showed that a much greater percentage of articles were documenting higher-level evidence of impact in the more recent decades.

Bennett’s Hierarchy was published in an issue of JOE dedicated to the evaluation of programming during 1975. This issue reported on the need to determine the value of Extension programming. Furthermore, the practice of program evaluation only recently gained importance during the 1960s. Therefore, it was expected that less reporting of evaluation results would be found in Cooperative Extension during the 1960s.

It was very interesting and meaningful to examine the differences of mean ranks for both level of Bennett’s Hierarchy and level of UWEX logic. The mean ranks have increased over time which indicated that more evaluative studies were reporting higher levels of evidence as the years progress. It takes time for a discipline to adapt, modify,
and apply a specific practice. When graphing the mean ranks the trend clearly showed that higher levels of evidence have been documented in more recent years.

A somewhat surprising finding involved the 1980s. The trend in the 1980s was somewhat flat meaning that little progress was achieved in terms of increasing the level of evidence collected. There are likely numerous reasons for this including evaluation capacity, knowledge, and skills; but it might also involve the priorities of society and the organization at that time. One possible explanation involves the farm crisis of the 1980s as Extension played a significant role in helping communities recover (Cartwright, Case, Gallaher, & Hathaway, 2002). Many Extension personnel were focused on addressing this problem, and were likely less concerned with the practice of program evaluation. The practice of program evaluation requires resources that aren’t readily available during times of crises or emergencies.

Objective 4: To identify and quantify the primary evaluation and data collection methods used to document evidence of impact.

Summary. The primary evaluation and data collection methods were categorized based on Diem’s (2003) examples of basic methods and tools used to measure impact. This study found that almost three-fourths used an evaluation method which asked (or surveyed) participants, almost one-third used an evaluation method which tested participants, and around 10% used an evaluation method which involved observation. For the actual data collection method, one-half (50.3%) of the evaluative studies used some form of a written questionnaire. The data collection method of pre-tests/post-tests
was used in 20.2% of the studies and interviews/testimonials/case studies/focus groups in 17.9%.

Discussion. The practice of program evaluation utilizes social science research methods. Survey research is one of the most important areas of measurement in applied social research (Trochim, 2006). It was not surprising that the majority surveyed participants. A little less than one-third utilized testing. Testing is important because it has the potential to provide much greater evidence of achievement than simply asking others. However, there is a drawback with testing when dealing with adult educational programs. Using a test is considered more of a pedagogical educational tool. Extension programming is typically voluntary and non-credit. More creative forms of testing are required to document adult knowledge gained as a result of programming. In certain situations, it might be beneficial to use weaker evidence of KASA change and then use more meaningful results to document evidence of practice change and end results.

Observation was used in the least amount of articles. In addition, reviewing information from other sources was coded as observation for this study. This means the use of direct observation was likely used in less than 10% of the articles. Observation requires time and attention, but could provide valuable results where conditions and the context are appropriate.

The written questionnaire was the predominant tool used to collect data. This would be expected as it is the most commonly used instrument for survey research. The use of pre-tests/post-tests was also frequently used. Testing is often the most appropriate method of documenting evidence of KASA change/short-term outcomes, which was the
most frequently reported level. It was somewhat surprising that interviews/testimonials/case studies/focus groups were the third most represented data collection method. This method requires more resources, but can be used with a very small group of people with results inferred to a larger group.

Objective 5: To determine if the program being evaluated was designed for reach (number of people interacted with) and/or richness (detail and depth of interaction).

Summary. The level of reach was fairly evenly distributed with about one-half of the articles coded as low and the other half coded as high. Fewer than 30% were coded as low richness with most considered as high richness. The concepts were also displayed as groupings. No programs were coded as low reach/low richness. Just over one-half of the articles were coded as low reach/high richness.

Discussion. The exploration of reach and richness and its relationship to documenting evidence of impact added a different aspect to this study. Should more resources of time, labor, and money be spent on more in-depth programming for fewer people? Should resources be targeted to as many people as possible using more mass media methods of delivery?

For the purposes of this study, the researcher decided to formulate the coding process so that it was required to select either high or low for each concept. These concepts were considered to be related (usually inversely related) and, thus, were analyzed together as groupings. There were no published articles regarding programs that were coded as low reach/low richness. This type of educational programming would
typically take place in one-on-one meetings or discussions which were not found as published evaluative studies in the JOE issues investigated. The majority of the articles (52.0%) were about programs that were considered to be low reach/high richness. Based on the description of the concepts developed by Evans and Wurster (2001), it was expected that programming with high richness would tend to have less people participating. The findings showed that around one-half of Extension programming involved in-depth programming. In-depth programming requires more individualized instruction and more two-way dialogue between instructors and participants.

Only 28.1% of programs were coded as high reach/low richness. These were programs that targeted the masses and provided limited depth of information. The high reach/high richness programs typically involved the evaluation of a program area, the system (CES), or an individual program which utilized an electronic means of delivery. The high reach/high richness programs consisted of 19.9% of the articles. The articles in which the object of evaluation was a program area or CES consisted of only 11.9% of all articles. Therefore, it is likely that there were a significant number of individual programs which utilized an electronic means of delivery that were coded as high reach/high richness.

The identification and use of these concepts originally reported by Evans and Wurster in 1997 has not been previously investigated with regards to Extension programming. This study contributed to the literature by providing Extension personnel an additional way of thinking about program development. An Extension educator who desires to make the greatest contribution must document evidence of end results/long-
term outcomes that affect societal conditions. The concepts of reach and richness provide an additional method for communicating the value of those end results.

High reach/high richness isn’t necessarily the goal of programming. The appropriate level of reach and richness depends on the subject matter, problem or issue, and context in which the programming is to be provided. A good way to address a problem or issue and meet an identified need would be to provide different forms of programming in which one is targeted towards reach and one is targeted towards richness. Those who benefit from high reach programming are potential clientele for future programming designed for high richness.

David King and Michael Boehlje (2000) used the concepts of “reach and richness” when discussing Extension programming to make the point that for Extension to remain relevant, personnel should become more focused on increasing reach (without losing richness) or private industry will attempt to fill the void. If unbiased Extension programming is not making a significant impact, then private industry (potentially biased) will become the predominant source of information and programming for current and future Extension clientele. The findings of this study showed that slightly more than one-half were low reach/high richness programming. This is likely the type of programming in which King and Boehlje were referring. Their point is very valid, but Extension should not increase reach at the expense of richness. To combat this problem, multifaceted programming for a specific problem or issue that involves a program focused on reach and a separate program focused on richness should work well. In
addition, the ability to design individual programs for both high reach and richness will likely increase as technology improves.

**Objective 6: To determine the geographic scope (county/local, multicounty regional, statewide, multistate regional, or national) of the programs evaluated.**

**Summary.** Most of the programs investigated were statewide in scope (51.0%) followed by multicounty regional (19.2%) and county/local (18.5%). There were fewer multistate regional (6.3%) and national (5.0%) programs evaluated.

**Discussion.** Extension programming is considered localized and community-based. However, the majority of the articles were reporting on statewide programs. This was likely because these are the ones that were most frequently submitted to the *Journal of Extension*. Nationwide, Extension personnel with tenure-track appointments tend to be specialists. Those with requirements for publications were likely more apt to evaluate programs and publish their findings. It was possible that the majority of program evaluation within Cooperative Extension occurred at the local level because this is where most of the programming takes place. The real aim of program evaluation is to gather information about the merit and worth of programming according to one’s criteria and then use that information to make future decisions. In accomplishing this aim, the findings might not be considered suitable for publication and the priorities of the agent/educator might be more directed towards program improvement/future programming and less on publishing results. This is understandable and often
appropriate, but sharing results helps others to learn and potentially improve their own methods.

As budgets become more restricted and technology advances, it is likely that there will be more multistate or national programming. Program evaluation of this form of programming is also likely to increase. In addition, county-specific programming will likely decrease if more state systems adopt regional structures similar to Minnesota.

**Objective 7: To explore the relationship between reach/richness, geographic scope, and the level of evaluation according to both Bennett’s Hierarchy and the logic model.**

*Summary.* There was a low negative relationship between Bennett’s Hierarchy/UWEX logic model and the level of reach. There was a very low positive relationship between Bennett’s Hierarchy/UWEX logic model and the level of richness. The direction of the relationship indicated that high reach programming tended to have less documentation of impact. Furthermore, the programs having higher richness tended to have more documentation of evidence of impact. The findings regarding geographic scope were very low.

*Discussion.* The relationships of reach and richness with level of evidence were low at best. The programs delivered with high richness tended to have higher levels of evidence documented, and the programs delivered with high reach tended to have lower levels of evidence. The lower magnitudes likely explain very little about the relationships. It is clear that more research is required to better understand the role of reach and richness with impact.
Based on the findings, geographic scope explained very little about the documentation of evidence of impact. However, the findings do show that geographic scope was a variable that was likely not very important when documenting evidence of impact. The variable likely has importance in program development and design, but evidence of impact can be collected and documented for a program of any scope.

Implications of the Conceptual Framework

The conceptual framework presented in Chapter 1 displayed the progression of the collection of evidence towards the ultimate goal of impact. The various levels of both Bennett’s Hierarchy and the components of the University of Wisconsin-Extension logic model format were displayed. The independent variables were considered to be Level-Bennett’s Hierarchy and Level-UWEX Logic Model. Each of the individual levels of evidence represented a different level of the independent variable. Impact was the dependent variable affected by the level of evidence. Programming might have tremendous impact; however, this cannot be known if evidence to document that impact is not collected.

The moderating variables which could potentially affect the overall quality or benefit of that impact (dependent variable) were considered to be reach, richness, and geographic scope. The findings of this study indicated a low association between reach/richness and the documentation of evidence of impact. Based on these findings, reach and richness does not appear to be a major moderating factor in reaching impact. However, additional research is necessary to further investigate the role of reach and
richness on achieving impact. Furthermore, the finding showed very low associations with geographic scope. Geographic scope does not appear to be a moderating variable for this conceptual framework.

Greater evidence of impact more strongly proves that impact has actually occurred. Level 1 (inputs & situation/priorities) provides no real evidence of impact. However, some people have reported the amount of money spent or number of people utilized to combat specific problems or issues with the intent to show impact. Level 2 (activities & inputs) also provides no real evidence of impact. Level 3 (people involvement & outputs-activities) provides little to no evidence of impact. It helps one describe the reach of involvement and/or what they did, but it is unknown if the activities contributed to anything of value. Level 4 (reactions & outputs-participation) provides somewhat weak evidence of impact. Participant’s reactions may indicate to program providers that a program might have some impact. If their reaction is very negative, it is unlikely that positive impact has occurred. Level 5 (KASA change & short-term outcomes) does provide some evidence of real impact. If knowledge, attitudes, skills and aspirations are changed; there is a real outcome resulting from programming. Level 6 (practice change & medium-term outcomes) is the level where more substantial impact is documented. At this level, program providers can see that KASA change has been applied and practice or behavior change has occurred. This evidence showed that programs were accomplishing something much more tangible. Level 7 (end results & long-term outcomes) is the ultimate goal and shows real impact. This evidence attempts to prove that societal conditions have changed as a result of the application of KASA;
thus, a positive change in social, economical, environmental, or civic conditions has occurred. Regardless of the reach, richness, or geographic scope; impact was achieved at level 7. The concepts of reach, richness, and geographic scope can be considered when planning for future programming to make additional, but different impact.

Recommendations

This section begins with recommendations for practice followed by recommendations to guide further research. The findings of this study suggest the following recommendations for Cooperative Extension professionals.

Recommendations for Practice.

1. When creating a logic model, Extension evaluators should understand and use Bennett’s terminology to plan some form of evaluation for each stage/level of their logic model. Bennett’s terminology was specifically created to describe different levels of evidence of impact. The components of the logic model were designed to help with program planning, development, and evaluation. How does one define or describe a short-term outcome or a medium-term outcome when communicating with others about their logic model? These components can appropriately be described as KASA change and practice/behavior change respectively.

Bennett’s seven levels are ideal for articulating and ensuring that evidence of impact is documented for each level. The following description provides specific
recommendations for collecting evidence of impact. The first three levels (inputs, activities, and people involvement) are clearly about the program. This information can be gathered from the program administrators or the Extension evaluator(s) themselves as they determined what resources were used, what occurred, and who or how many attended. The fourth level is reactions which can be documented by simply asking the participants. A more sophisticated version of the “smile sheet” passed out at the end of a session or program can be used to collect reactionary data. The fifth level is KASA change. This can easily be accomplished by using a pretest/posttest or some form of testing (i.e., post-then-pre). This instrument should be something more complex and rigorous than the basic “smile sheet” used to collect reactionary data. The sixth level is practice change. Practice change can be measured by using a survey questionnaire after a significant amount of time has passed since the program ended. Participants must have sufficient time to apply and practice their new knowledge before one can determine if adoption or behavior change has actually occurred. The seventh and final level is end results. This requires a change in social, economic, civic, or environmental conditions. The evaluator must wait until sufficient time has passed following practice change in order to determine the effect on society.

A potential rule of thumb for evaluating Extension programming is as follows: (1) collect reaction data at the end of each session or at the end of the program; (2) collect KASA change data with a pretest at the beginning and posttest at the end of each session, or the program; (3) collect practice change data six months to two years
following program completion; and (4) collect data regarding the end results three to five years following program completion.

2. More documentation of end results or long-term outcomes is needed when evaluating programs. Currently, only 5.6% of the published outcome studies from the applicable years attempted to document evidence of long-term outcomes. End results are the broader outcomes, effects, and benefits to the community or society such as increased profits, better health, or cleaner environment. They help accomplish the overall mission of CES by positively affecting social, economical, civic, and environmental conditions for society. A change in behavior and practice demonstrates that programming has made a substantial impact, but what does it mean for society? The application of knowledge that results in practice change could potentially have a negative long-term impact on society. Data should be collected to determine the positive end results and long-term outcomes as this provides the greatest evidence of real or actual impact. This is the kind of information that administrators and legislators truly need in order to ensure that public resources are being properly utilized and directed to the most beneficial areas.

3. Reach and richness are important concepts to understand and consider when creating and planning programming. A program with tremendous value for a few individuals is less likely to achieve impactful end results on societal conditions. Even if the program documented evidence of substantial impact for those few participants, it is up to the Extension professional to see that the information eventually reaches a much larger audience. On the other extreme, if a program reaches a tremendous amount of
people with superficial information, it is less likely to achieve impactful end results on societal conditions. However, if Extension helps a single individual who is now in a position to create a significant number of jobs for others as a result of programming, this is a benefit to societal conditions that should be documented for accountability. Extension Educators would be well served to balance their programming efforts to ensure that they are accomplishing both. For example, two programs could be developed to address a single issue or need in which one program focuses on reach and the other focuses on richness. In the future, the ability to design individual programs for both high reach and richness will likely increase as technology continues to improve.

4. Evaluation efforts should be focused on evaluating individual programs (instead of a state’s system/program area and/or CES) in order to collect more valid data and more efficiently utilize resources. If enough individual programs are evaluated, then results can be inferred to a state’s system/program area and/or CES. Those large-scale evaluations of CES or an entire state’s program area are difficult to administer and require tremendous resources to collect valid results. The one exception for program area would be 4-H Youth Development. Because of the fact that this program area has enrolled membership, large-scale evaluations tend to be more feasible and valid. The Cooperative Extension System is extremely large and complex; however, when evaluation is conducted on individual programs by people working closely with that specific program, the almost impossible task becomes achievable. Each can do their part by assuring that their own individual programs are evaluated either by
themselves or an external entity. As the findings of this study showed, those evaluations of the whole system (CES) tended to document lower levels of evidence.

5. The creation of a statewide template or reporting system would allow Extension professionals to enter their individual program results into a database in which evaluative conclusions could be generated for the larger organization. For example, a template was created for the statewide ServSafe™ program (Boyer, Benson, Boyd, Forrester, Franz, Gehrt, Pelland, & Roan, 2009). Extension professionals could enter their results from individual programs for each of the seven levels of evidence into a similar database. In addition, background information about the program could be entered to examine the findings across specific program areas, program types, or across CES as a whole. This is similar to how the object of evaluation was used in this study. Individual programs for clientele are somewhat different than individual programs for personnel (in-service). Individual tools such as newsletters, bulletins, web sites, television/radio programs are somewhat different than traditional face-to-face programs. The results from various types or forms of programming may have dissimilar meanings and impact society differently. The impact on society from an Extension in-service is much more indirect than the impact resulting from a program administered for clientele. Generating statewide (or nationwide) impact data from Extension programs can be challenging (Boyer et al., 2009). The template or database system of reporting results provides a more feasible way to evaluate an entire program area or the U.S./State Cooperative Extension System.
6. Extension should seek external funding by reporting on evidence of previous impacts and the future impact that will result from additional funding instead of only focusing on the need. Funders will still want to know why their money is needed and the importance of that need. In addition, they will want to gain some understanding of how one will go about spending the money to meet that need. However, the use of impact can greatly enhance the proof that funding will make a positive difference. A description of similar or related programming that used “X” number of dollars, people, and time resulted in “X” impact based on “X” documented evidence provides proof that funding has truly worked successfully in the past. In addition, the expected impact and evidence that will be collected as proof is also extremely important. Hughes and Ledbetter (2009) made the following recommendations for Extension agents: understand the relevance to the community of their programming, ensure reliability to conduct successful programming, and demonstrate ability to communicate impact to stakeholders. The terminology of Bennett’s Hierarchy provides a real method to communicate different levels of evidence of impact.

**Recommendations for Future Research.**

1. Additional research is required to determine the quality (worth or merit) of evaluative studies published by Extension professionals. The present study simply identified the level of evidence of impact collected in evaluative outcome studies that were published in the *Journal of Extension* during the years reviewed. There were no judgments or measures as to the quality of the methods and procedures used or the
significance of the findings. Research investigating the quality of the reported
program evaluations would be very similar to conducting a metaevaluation which is
an evaluation of a program evaluation. Criteria should be developed to determine the
quality of the methods used to collect the data. For example, an interview that
provided the reaction of an Extension agent/educator would be of much less value
than an experimental design with a comparison group. In addition to the
methodology used, the value of the results can be determined. Findings reporting that
participants on average increased their scores on the post-test by 5% shows evidence
of KASA change, but does it show meaningful evidence of KASA change? A coding
system or criteria is needed to determine the quality of the methods and the
significance of the findings. In addition, there was no attempt to look at program
evaluations reported in other sources.

2. Additional research is needed to better understand the relationship of reach and
richness with impact. The concepts are important because even though a high level
of impact might be documented by behavior change and the application of
knowledge, the overall benefit of that impact depends on the significance of the
number of people reached and the degree of richness. It would be helpful to develop
operational definitions for both concepts that allow the researcher to use various
coding levels for each concept. This exploratory study simply categorized the
concepts as either high or low.

3. More investigation is required to investigate the documentation of evidence regarding
end results and long-term outcomes. The results of this study have shown that the
highest level of documentation of impact was not represented very well in the *Journal of Extension*. In addition, this study did not determine the worth or merit of those outcomes; therefore, it is possible that the percentage of studies documenting significant end results with real impact was even lower than 5.6%. It is important to focus on those published studies that do provide evidence of impact on societal conditions. By determining the methods and tools used to collect evidence of end results/long-term outcomes, other Extension evaluators will likely be more apt to collect this evidence.

**Conclusions**

This study was initiated to explore and describe the use of program evaluation and the documentation of evidence of impact in Extension programming. Qualitative interpretations, based on reviewing such a large portion of the *Journal of Extension*, were that Extension has traditionally had a very firm grasp on identifying problems, issues, and needs. There appeared to be many articles published regarding the assessment of needs. However, only about one-fourth (302/1,446) of the articles published, during the applicable years (1965-69, 1975-79, 1985-89, 1995-99, & 2005-09), examined programmatic outcome data to determine if those needs were being met.

The number of evaluative studies with outcome data has increased going from 16 articles in the 1960s to 150 articles in the 2000s. The findings indicated that 88.5% of the articles documented evidence above the level of participation (levels 4-7), almost two-thirds (62.6%) were measuring outcomes (levels 5-7), about one-third (32.8%) were
measuring evidence of substantial impact (levels 6-7), and 5.6% documented “real or actual” impact (level 7).

Reach and richness are important concepts to keep in mind when creating impactful programming to ensure an adequate balance of information dissemination. More research is needed to learn more about the role of these concepts in regards to impact. However, the overall impression from reading the articles was that the authors made no distinction in programming based on the reach of participation or the richness of content. Although, authors often reported numbers of participation, but there was seldom any information regarding the targeted audience that chose not to participate.

The documentation of evidence of impact is essential to help ensure that funding exists in the future. The time spent documenting behavior change may result in fewer programs but increased impact of educational efforts (Clements, 1999). All stakeholders including both program providers and recipients need to be aware of the impact of Extension programming. There must be evidence provided to document or prove impact because simply telling someone or assuming worth will be self-evident is very likely not enough. Because the clientele of the Cooperative Extension Service are the general public, all citizens are considered to be stakeholders. Funding comes from federal, state, local, and extramural (grants, contract, gifts, user fees) sources. Legislators, grantors, donors, and users will determine who receives future funding. Programs are publically funded because voters/campaign contributors want them to be funded (Martin, 2002). A recent random sample (32% response rate) of Ohio residents reported that only one-fifth
of respondents were familiar with Extension (Loibl, C., Diekmann, F., & Batte, M. T., 2010). What portion of U.S. citizens are likely aware of its impact?

It is highly likely that Extension programs make a tremendous positive difference in the lives of individuals and their communities and Extension greatly benefits society as a whole, but more program evaluation with evidence of higher-level impact is needed to support this argument. It is also likely that funding becomes more limited and much more competitive moving into the future. Those organizations who have real documented evidence of their worth and merit will be the ones that receive the funding and continue to exist in the future.
LIST OF REFERENCES


135


Homestead Act of 1862, Public Law 37-64, 05/20/1862; Record Group 11; General Records of the United States Government; National Archives.


APPENDIX A
RUBRIC

This study uses a form of quantitative content analysis. Due to the qualitative nature of the data collection process, Stage I should be completed in its entirety prior to beginning Stage II in order to help maintain consistency in the selection of articles.

The instrument consists of a Microsoft® Office Word® document to use as a questionnaire and a Microsoft® Office Excel® spreadsheet to use for data entry.

<table>
<thead>
<tr>
<th>INSTRUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ Publication Year _______ Volume _______ Number _______</td>
</tr>
<tr>
<td>➢ Lead author’s last name: ____________________________</td>
</tr>
<tr>
<td>➢ Category of JOE in which the article is listed:</td>
</tr>
<tr>
<td>1 Feature Articles</td>
</tr>
<tr>
<td>2 Research in Brief</td>
</tr>
</tbody>
</table>

The first portion of the instrument consists of organizational data (similar to demographic data) which was pre-entered into the spreadsheet.

Stage I – Selection of Articles (screening abstracts)

1) After screening the abstract, should the article be selected for this study?
   0 No
   1 Yes
Question #1 of the questionnaire involves the selection of articles and should be completed during Stage I by screening abstracts/articles. During the earlier issues of JOE, a true abstract was often not included. Generally, the first paragraph of the actual article can be used in place of a true abstract in these particular articles; however, further reading or scanning during the screening process is sometimes required (also includes very short or incomplete abstracts).

*How to select an article while screening the abstract? (All must be answered as “yes”.)*

1. Does the article involve the systematic acquisition and assessment of information to provide useful feedback about some object?
2. Does the article involve the acquisition and use of programmatic information?
3. Can the article be considered as a program evaluation meaning was the purpose to determine the value (worth or merit) of something? *(How good is something?)*
4. Can the type of program evaluation be considered as an outcome study in that an intervention (programming) has already occurred and results (outcome data) were collected about the programming?
5. Can an object of the evaluation be identified and does it belong to the U.S. Cooperative Extension System (CES)? *(It is required that CES conducted the intervention.)*

✓ Identifying and thinking about the object of the evaluation is necessary to determine if an article should be selected for this study. The evaluation object is defined as whatever is being evaluated. Surveying program participants is a method frequently utilized to evaluate a program; thus, the primary object of the evaluation is the actual program being evaluated not the respondents (i.e., participants, volunteers, personnel).

✓ Some articles claim to be program evaluations or evaluative studies while others are simply research studies (i.e., exploratory/descriptive, explain/predict, experimental). Instead of determining if an article is research or evaluation, a judgment must be made as to whether the findings could be considered as evaluative data resulting from programming. In order to be evaluative data, an object of the evaluation must be identified.
To be selected for the purposes of this study, the object of the evaluation must be some form of educational programming; either an individual program, an educational tool, or a program system (program area or the U.S./State CES system). The educational programming must be a part of the U.S. Cooperative Extension System meaning that it is essential that CES conducted the intervention (or programming). Partnering with other organizations or entities is acceptable. When the evaluation object is an educational tool, CES must have created, modified, or utilized the tool for Extension programming. Extension personnel occasionally evaluate non-Extension programs or objects. These types of articles should not be selected for purposes of this study.

How to eliminate (or not select) an article while screening the abstract?

1. No data or information was collected.
2. The data or information cannot be considered as programmatic information.
3. The article is clearly a research study in which the results could not be used for evaluation (determining the value [worth or merit] of an object).
4. The article was not reporting the results of programming (outcome study). (It was solely a needs assessment or process study [monitoring study]).
5. The object is not a part of the U.S. Cooperative Extension System (when the article was written). For programming, there was no intervention by CES. The educational tools were not created, modified, or utilized by CES.

Articles regarding the assessment of needs were not selected if the previous criteria were not satisfied. Needs assessments were only selected if the assessment was somehow incorporated into an outcome evaluation (determining value) being reported in the article. For example, a current program assesses what else is needed while collecting evaluative information regarding what has already been delivered.

Process or monitoring studies were also not selected if the previous criteria were not satisfied. Process or monitoring studies were only selected if they also included evaluative results (outcome data) about existing programming.

Articles can also be eliminated after selection (abstract screening process) if it’s eventually determined that the respective article did not meet the above criteria.
Stage II – Completion of the Instrument (data collection)

✓ The remainder of the questionnaire (questions 2 – 12) is completed during Stage II by reading the selected articles.

✓ It is important during Stage II to discard any articles selected during the screening process in Stage I which are found to not meet the criteria required for purposes of this study. A typical reason for discarding an article would result from reading the introduction or a portion of the article and determining that the program being evaluated was not an Extension program (U.S. CES). Another reason would be that the data collected was not considered to be results (outcome data) as no previous intervention (programming) had actually taken place.

2) What is the lead author’s current home state listed in the article? ________________

Please select the state listed in the description of the lead author or take directly from their title or e-mail address. Do not use the state in which the article reports or data represents.

3) Within which category does the reported occupation of the lead author best fit? (max of two)
   1 Extension Professional (i.e., Agent/Educator)
   2 Extension Professional (i.e., Specialist)
   3 Extension Professional (i.e., Director, Leader, Head, Administrator)
   4 Extension Staff (i.e., Extension Associate; Program Assistant/Coordinator/Manager)
   5 University Faculty (i.e., Assistant Professor/Associate Professor/Professor)
   6 University Student (i.e., Graduate Assistant/Associate)
   7 other
   8 unknown

Please select the occupation(s) listed in the description of the lead author or take directly from their title. Do not infer or retrieve their occupation based on the article.
This is inferred from reading the article. Please choose all that apply. For those articles reporting on the U.S. CES in general, please select all program areas.

5) What is the object of the evaluation?
   1 Individual Programs for Extension Clientele/Volunteers
      (evaluatees such as: participants or recipients)
   2 Educational Tools for Extension Clientele/Volunteers
      (evaluands such as: newsletters, bulletins, web sites, television or radio programs)
   3 Individual Programs for Extension Personnel (e.g., in-service / training programs)
   4 Program Area (i.e., Agr & NR, 4-H, FCS, CED)
   5 U.S./State Cooperative Extension System

The evaluation object is defined as whatever is being evaluated. Surveying program participants is a method frequently utilized to evaluate a program; thus, the primary object of the evaluation is the actual program being evaluated not the respondents (i.e., participants, volunteers, personnel).

These five objects represent the different forms of CES programming. Articles evaluating these five objects should be selected for purposes of this study. Other objects of evaluation do not meet the purposes of this study (i.e., non-Extension programs; instructional methods; delivery methods; personnel/volunteer traits, skills, or knowledge). An example where careful attention must be used involves articles that also report
programmatic results (outcome data) of individual programs (‘1’) while their purpose for writing the article and primary evaluation object was actually delivery method. These articles are selected because they also include the evaluation object of individual programs (‘1’).

It is important to note that the program area of 4-H Youth Development is somewhat different from the other three program areas. This is because 4-H Youth Development has enrolled members, and is much more likely to conduct evaluative studies on an entire state’s program area or the 4-H program in general (‘4’).

<table>
<thead>
<tr>
<th>6) What is the highest level of evidence according to Bennett’s Hierarchy documented in this article?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (inputs)</td>
</tr>
<tr>
<td>2 (activities)</td>
</tr>
<tr>
<td>3 (people involvement)</td>
</tr>
<tr>
<td>4 (reactions)</td>
</tr>
<tr>
<td>5 (KASA change)</td>
</tr>
<tr>
<td>6 (practice change)</td>
</tr>
<tr>
<td>7 (end results)</td>
</tr>
</tbody>
</table>

Inputs include the resources invested by Extension such as time, money, and staff. Activities are meetings, sessions, workshops, events, etc. that Extension personnel provide and coordinate. People involvement is reporting the number of people who attended and other information about who participated (demographic data). Reactions are what participants thought of the program, its organization, its leader, etc. Reactions can be both good and/or bad. KASA change is a measurement of changes in knowledge, attitudes, skills, and/or aspirations. Practice change is considered to be behavior change
or application and occurs only if the changes in knowledge, attitudes, skills, and/or aspirations are actually applied. End results are the broader outcomes, effects, and benefits to the community or society such as increased profits, better health, or cleaner environment. End results help accomplish the overall mission of CES such as positively affecting social, economical, civic, and environmental conditions for society.

7) What is the highest level of evidence according to the University of Wisconsin-Extension logic model format documented in this article?

1 (situation/priorities)
2 (inputs)
3 (outputs-activities)
4 (outputs-participation)
5 (short-term outcomes)
6 (medium-term outcomes)
7 (long-term outcomes)

The situation is the originating program or issue set within a complex of sociopolitical, environmental, and economic circumstances. Priorities are the types of information gathered from needs assessments, but may also be included when reporting the results or outcomes of programming. Typically corresponds to Bennett’s category of inputs. The inputs are the resources and contributions that go into or are invested into the program. Also corresponds to Bennett’s category of inputs. Outputs are what the program does and who it reaches such as activities, services, events, products, and people reached. Bennett’s category of people involvement will typically correspond with outputs-participation. Bennett’s category of reactions will typically fall into either outputs-activities or outputs-participation. Bennett’s category of activities will typically correspond with outputs-activities. Outcomes are the results such as the value or changes
for individuals, families, groups, agencies, businesses, communities, and systems. Short-term outcomes are immediate and include changes in learning (i.e., awareness, knowledge, skills, attitudes, opinions, intent). Always should be the same as Bennett’s category of KASA change. Medium-term outcomes are intermediate and include changes in action (i.e., behaviors, decision-making, actions). Always should be the same as Bennett’s category of practice change. Long-term outcomes (often referred to as impact) are the final outcomes which include changes in conditions (i.e., social, economic, civic, environmental). Always should be the same as Bennett’s category of end results.

8) What is the primary evaluation method(s) utilized in this article? (maximum of two)
   1 (ask them-survey)
   2 (test them-simple experimental designs)
   3 (observe them-observations)

The evaluation methods are somewhat self-explanatory. If the evaluator/research is reviewing information from other sources, then they are observing.

9) What is the primary data collection method(s) utilized in this article? (maximum of two)
   1 (written questionnaires)
   2 (longitudinal studies)
   3 (interviews/testimonials/case studies/focus groups)
   4 (pre-test/post-test)
   5 (post-test with control group comparison [i.e., both true- & quasi- experimental])
   6 (direct observation)
   7 (reviewing information from other sources)
   8 (online survey)
   9 (retrospective pretest / post-then-pre)
   10 (other)
The data collection methods are also somewhat self-explanatory, but do require some prior knowledge regarding research methods.

10) What is the level of reach of the program represented in this article?
   1 (low)
   2 (high)

Reach is the amount of individuals you can touch and influence (with information). For example, programming delivered through use of mass media would usually be considered as having high reach while a short course would usually be considered as having low reach.

Criteria for high reach programs were those that generally (1) were created and targeted to reach an extremely large audience, (2) did not require any formal registration or attendance, and/or (3) did not contain any evidence or record of participation by any single individual. An example of high reach programs would be those providing information through mass media.

Criteria for low reach programs were the exact opposite and those that generally (1) were created and targeted to reach a smaller and more well-defined audience, (2) required formal registration and/or attendance, and/or (3) contained evidence of participation and attendance.
Richness is the detail and depth of the interaction (with information). For example, programs that require participants to attend or participate multiple times allowing them the opportunity to practice and apply their new knowledge and receive feedback and more individualized instruction throughout the program’s duration are typically considered as having high richness. Generally, one shot programs have low richness.

Criteria for high richness consisted of programs that generally had (1) high bandwidth in which very in-depth information was moved from sender to receiver, (2) high customization in which the program was specifically tailored to a specific audience, (3) high interactivity in which there was a great deal of exchange between program providers (e.g., instructors, administrators) and participants, (4) high reliability in which participants tended to trust program providers, (5) high security in which sensitive information could be shared, and/or (6) high currency in which information exchange (feedback) was immediate.

Criteria for low richness were considered the exact opposite with programs that generally had (1) low bandwidth in which very general information was provided, (2) low customization in which the program was not specifically targeted to a specific audience, (3) low interactivity in which there was little to no exchange between program providers.
(e.g., instructors, administrators) and participants, (4) low reliability with less trust as participants and program providers were likely strangers, (5) low security in which sensitive information was unlikely to be shared, and/or (6) low currency in which information exchange (feedback) was slower.

*** Reach is usually negatively correlated with richness unless there was a digital network or electronic means utilized for the delivery of the information (e.g., web sites, CD ROM, television, radio). However, more complex Extension programming (especially when the evaluation object is program area or U.S./State CES) can and often does both and should be coded accordingly (positively correlated).

Evans and Wurster (2000) have stated, “to the extent that information is embedded in physical modes of delivery, a basic law governs its economics: there is a universal trade-off between richness and reach. But unbundle information from its physical carrier, and the richness/reach trade-off can blow up.” Stated another way, “when information is carried by things – by a salesperson or by a piece of direct mail, for example – it goes where the things go and no further. It is constrained to follow the linear flow of the physical value chain. But once everyone is connected electronically, information can travel by itself. The traditional link between the flow of product related information and the flow of the product itself, between the economics of information and the economics of things, can be broken (Evans & Wurster, 1997).”
12) What is the geographic scope of the program represented in this article?
   1 (county [local])
   2 (multicounty regional)
   3 (statewide)
   4 (multistate regional)
   5 (national)

The geographic scope is the geographic region/area or location of the clientele/audience targeted for the programming represented in the article/study. The coding for geographic scope of programming has no bearing on the reach and richness. Both local and national programming can have high reach and low richness, or vice versa.
APPENDIX B

INSTRUMENT – QUESTIONNAIRE

- Publication Year ________ Volume ________ Number ________
- Lead author’s last name: ______________________
- Category of JOE in which the article is listed:
  1 Feature Articles
  2 Research in Brief

1) After screening the abstract, should the article be selected for this study?
   0 No
   1 Yes

2) What is the lead author’s current home state listed in the article?
   ______________________

3) Within which category does the reported occupation of the lead author best fit? (max of two)
   1 Extension Professional (i.e., Agent/Educator)
   2 Extension Professional (i.e., Specialist)
   3 Extension Professional (i.e., Director, Leader, Head, Administrator)
   4 Extension Staff (i.e., Extension Associate; Program Assistant/Coordinator/Manager)
   5 University Faculty (i.e., Assistant Professor/Associate Professor/Professor)
   6 University Student (i.e., Graduate Assistant/Associate)
   7 other
   8 unknown
4) What is the predominant major program area(s) represented in this article?

1 (Agriculture and Natural Resources)
2 (4-H Youth Development)
3 (Family and Consumer Sciences)
4 (Community and Economic Development)
5 (combination of program areas) i.e., ________________________________

5) What is the object of the evaluation?

1 Individual Programs for Extension Clientele/Volunteers
   (evaluatees such as: participants or recipients)
2 Educational Tools for Extension Clientele/Volunteers
   (evaluands such as: newsletters, bulletins, web sites, television or radio
    programs)
3 Individual Programs for Extension Personnel (e.g., in-service / training)
4 Program Area (i.e., Agr & NR, 4-H, FCS, CED)
5 U.S./State Cooperative Extension System

6) What is the highest level of evidence according to Bennett’s Hierarchy documented
   in this article?

1 (inputs)
2 (activities)
3 (people involvement)
4 (reactions)
5 (KASA change)
6 (practice change)
7 (end results)

7) What is the highest level of evidence according to the University of Wisconsin-
   Extension logic model format documented in this article?

1 (situation/priorities)
2 (inputs)
3 (outputs-activities)
4 (outputs-participation)
5 (short-term outcomes)
6 (medium-term outcomes)
7 (long-term outcomes)
8) What is the primary evaluation method(s) utilized in this article? (maximum of two)

1 (ask them-survey)
2 (test them-simple experimental designs)
3 (observe them-observations)

9) What is the primary data collection method(s) utilized in this article? (maximum of two)

1 (written questionnaires)
2 (longitudinal studies)
3 (interviews/testimonials/case studies/focus groups)
4 (pre-test/post-test)
5 (post-test with control group comparison [i.e., both true- & quasi-experimental])
6 (direct observation)
7 (reviewing information from other sources)
8 (online survey)
9 (retrospective pretest / post-then-pre)
10 (other)

10) What is the level of reach of the program represented in this article?

1 (low)
2 (high)

11) What is the level of richness of the program represented in this article?

1 (low)
2 (high)

12) What is the geographic scope of the program represented in this article?

1 (county [local])
2 (multicounty regional)
3 (statewide)
4 (multistate regional)
5 (national)
APPENDIX C

INSTRUMENT – DATA ENTRY FORM EXAMPLE

<table>
<thead>
<tr>
<th>Year</th>
<th>Vendor</th>
<th>Vendor Code</th>
<th>Vendor Name</th>
<th>Address Line 1</th>
<th>Address Line 2</th>
<th>City</th>
<th>State</th>
<th>ZIP Code</th>
<th>Contact Person</th>
<th>Contact Phone</th>
<th>Contact Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>1</td>
<td>401</td>
<td>C. Thomas</td>
<td>123 Street</td>
<td>456 Avenue</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>John Doe</td>
<td>555-1234</td>
<td><a href="mailto:john.doe@email.com">john.doe@email.com</a></td>
</tr>
<tr>
<td>1996</td>
<td>2</td>
<td>402</td>
<td>Miller</td>
<td>789 Jackson</td>
<td>012 Main St</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>Jane Smith</td>
<td>555-5678</td>
<td><a href="mailto:jane.smith@email.com">jane.smith@email.com</a></td>
</tr>
<tr>
<td>1997</td>
<td>3</td>
<td>403</td>
<td>Smith</td>
<td>321 Maple</td>
<td>987 State St</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>John Doe</td>
<td>555-9876</td>
<td><a href="mailto:john.doe@email.com">john.doe@email.com</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Vendor</th>
<th>Vendor Code</th>
<th>Vendor Name</th>
<th>Address Line 1</th>
<th>Address Line 2</th>
<th>City</th>
<th>State</th>
<th>ZIP Code</th>
<th>Contact Person</th>
<th>Contact Phone</th>
<th>Contact Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>4</td>
<td>404</td>
<td>Brown</td>
<td>4567 Pine St</td>
<td>890 Oak Ave</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>Mary Jones</td>
<td>555-5555</td>
<td><a href="mailto:mary.jones@email.com">mary.jones@email.com</a></td>
</tr>
<tr>
<td>1999</td>
<td>5</td>
<td>405</td>
<td>White</td>
<td>234 Cedar Rd</td>
<td>123 Elm St</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>Susan Williams</td>
<td>555-6666</td>
<td><a href="mailto:susan.williams@email.com">susan.williams@email.com</a></td>
</tr>
<tr>
<td>2000</td>
<td>6</td>
<td>406</td>
<td>Green</td>
<td>789 Oak Blvd</td>
<td>456 Maple Ave</td>
<td>City</td>
<td>State</td>
<td>ZIP Code</td>
<td>John Smith</td>
<td>555-7777</td>
<td><a href="mailto:john.smith@email.com">john.smith@email.com</a></td>
</tr>
</tbody>
</table>

Note: The above table is a sample of the data entry form example. The actual form may have more fields and data entries.