4-2016

Rubrics as a Tool for Reaching Explicitly Evaluative Conclusions: Implications for Evaluation Theory, Training, and Practice

Krystin S. R. Martens
Western Michigan University

RUBRICS AS A TOOL FOR REACHING EXPLICITLY EVALUATIVE CONCLUSIONS: IMPLICATIONS FOR EVALUATION THEORY, TRAINING, AND PRACTICE

by

Krystin S. R. Martens

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation
Western Michigan University
April 2016

Doctoral Committee:

Lori Wingate, Ph.D., Chair
Michael Scriven, Ph.D., Director
E. Jane Davidson, Ph.D.
Marianne Di Pierro, Ph.D.

RUBRICS AS A TOOL FOR REACHING EXPLICITLY EVALUATIVE CONCLUSIONS: IMPLICATIONS FOR EVALUATION THEORY, TRAINING, AND PRACTICE

Krystin S. R. Martens, Ph.D.

Western Michigan University, 2016

Evaluation is the systematic determination of merit, worth, or significance. A core professional evaluator competency is to provide transparent and explicit evaluative conclusions. Yet, "understanding the reasoning process to establish evaluative conclusions drawn in practice has to be the field's greatest unmet challenge" (Fournier, 1995, p. 1). This three-article dissertation studies rubrics as a tool that can assist in meeting the stated challenge when used by program evaluators to reach explicitly evaluative conclusions.

Study 1 provides an account of the history and etymology of the term *rubric* and provides an analysis of peer-reviewed program evaluation literature, specific to the extent and how rubrics are portrayed in program evaluation. The results of the literature review produced few examples of the use of rubrics in program evaluation to reach explicitly evaluative conclusions.

Study 2 investigates the ways that evaluators use rubrics as evaluation-specific tools in program evaluation, and explicates how they learned to do so. Study 2 presents results of interviews with practitioners identified in Studies 1 and 2 as users

of rubrics to reach evaluative conclusions. Interviewees found rubrics to be useful in multiple ways, including reaching explicitly evaluative conclusions, but they rarely publish their experiences in the peer-reviewed literature. Guidance about this practice is, instead, typically shared through mentoring.

Study 3 fills a major gap in the program evaluation literature by explaining how the form (characteristics and configuration) and function (the natural purpose) of rubrics exemplify the core logic and nature of evaluation. This explanation can also promote movement toward a shared language that will enable theorists, researchers, trainers, and practitioners, who often hail from disparate academic backgrounds, to more effectively further theory, training, and practice of rubric use by program evaluators to make evaluative reasoning explicit.

Fournier, D. (1995). Editor's notes. *New Directions for Evaluation, 1995*(68), 1-4.

ACKNOWLEDGMENTS

My journey to degree completion would not have been possible without the support of many people. I would like to start by thanking the members of my dissertation committee for their time and support. First and foremost, I would like to thank Dr. Michael Scriven who was my advisor from the day I began my Ph.D. journey. He chaired my program committee and was chair of my dissertation committee until his departure from WMU forced an administrative change, and he was named the director of the committee. No matter the title, his teachings, advice, and guidance have greatly influenced my theory and practice. I am deeply moved by the generosity of Dr. Lori Wingate, who stepped in to chair my dissertation committee at a time when I was struggling to conceptualize a project that would both make a significant contribution to the field, yet be practical and feasible. Her feedback and guidance were instrumental in shepherding me through the process and substantially improving my work. She devoted a noteworthy amount of time to the effort, for which I am extremely grateful. I was honored to have Dr. Jane Davidson as an external committee member, as I have long admired her work. She provided valuable insights into the theory and practice that underpin this research and continually challenged my thinking along the way. Finally, I was privileged that Dr. Marianne Di Pierro was willing to join the committee. Her vast experience and intellect were instrumental to the process.

I would also like to thank the members of my program committee. As noted above, Dr. Michael Scriven, as chair, provided immeasurable evaluation expertise; Dr. Brooks Applegate provided extensive expertise in research design; and Dr. Carol

Weideman was a valued content expert in my cognate area of health and wellness. Their early commitment to the integrity of my program of study set the foundation for my successful completion of this dissertation research.

There are a host of other esteemed mentors, colleagues, and evaluation practitioners who generously gave their time to assist me. I have a special place in my heart for Mary Ramlow who made sure I stayed on track administratively and provided well-timed pep talks. Among those who contributed to early conceptualizations of my thesis, I thank Dr. Daniela Schröter, Dr. Chris Coryn, Dr. Lisa Russell, and BA Laris. I give my heartfelt gratitude to my colleagues at the University of Melbourne, especially Dr. Amy Gullickson and Dr. Janet Clinton, who helped me to balance my work-life and degree completion activities by providing the day to day support that I needed during the home stretch.

I am grateful to so many friends and colleagues who I have met and worked with along the way—you know who you are. Finally, I am indebted to all the friends and family who dealt with being leaned on, with little reciprocation, for much longer that should have been tolerated.

Lastly, my husband, Paul, provided stalwart support. When life kept happening *to* us—as we kept getting knocked down by issues outside of our control—he would pick me up, dust me off, and push me right back out there with an enduring (and somewhat inflated) belief in my abilities. At those times when I felt defeated, I just needed to look into the eyes of someone who believed.

Krystin S. R. Martens

TABLE OF CONTENTS

Table of Contents—Continued

Table of Contents—Continued

Table of Contents—Continued

LIST OF TABLES

LIST OF FIGURES

List of Figures—continued

I. INTRODUCTION

The Problem

Evaluation is the systematic determination of merit, worth, or significance (AEA, 2014). If the deliverable of professional evaluations are reports that include conclusions and recommendations, logically, whether or not evaluation conclusions are infused with evaluative terms (e.g., good, effective, worthwhile), the fact that conclusions are specified and recommendations are given denotes that at some level, judgments have been made (e.g., suggestions for improvement or about whether to continue, expand, reduce, or reinvent are given). To make such judgments, one has to have some set of standards to guide reasoning (Davidson, 2005; Scriven, 1991). Arens (2005) found in her study of Outstanding Evaluations, as awarded by the American Evaluation Association, that the conclusions and recommendations of the five studies she examined relied heavily on implicit criteria. If the interpretation process of award-winning evaluations is not transparent and explicit, others are unable to learn from these exemplary practitioners. It also hinders replicability, which diminishes validity and feeds into the pervasive notion that evaluation is subjective. One tool that can aid evaluators in explicitly conveying the standards that guide interpretation in evaluation is a rubric.

Rubrics are well-established tools used in a variety of educational settings, such as student assessment, teacher performance, and curriculum evaluation. Michael Scriven (1991) characterizes evaluation of student learning as performance evaluation, evaluation of staff performance as personnel evaluation, and evaluation of curriculum as product

evaluation. Rubrics are frequently used to make judgments for these purposes. There are many examples of evaluations conducted in education settings in the peer-reviewed literature that contain measures obtained by using rubrics to assess individual performance, personnel, and/or products, but few examples of evaluations that use rubrics to judge *program* performance. This research study investigates the use of rubrics as tools to combine evidence and values to draw evaluative conclusions in program evaluation.

Defined by the Oxford Dictionary (2016) *value* is "the regard that something is held to deserve; the importance, worth, or usefulness of something." In evaluation, values are context-dependent, but remain what is thought to be important or good—for example, democratic values (House & Howe, 1999). However, in evaluation values are rationally determined, and not based on subjective preferences, beliefs, or interests (House & Howe, 1999). Instead, in evaluation values are drawn from sources such as a needs assessment, literature, professional standards, or a logical analysis of the function of the evaluand (Davidson, 2005; Scriven, 1991). Values in evaluation are operationalized through criteria.

This dissertation is comprised by three separate studies. The first two studies address the gap in the literature surrounding rubric development and use in evaluation. Study 1 investigates how rubrics are presented in the peer-reviewed program evaluation literature. Study 2 investigates how evaluators who are using rubrics in an evaluation-specific way (by using rubrics to combine evidence and value) go about this practice and how they learned to do so. The first two studies inform the third study. Study 3 helps fill a major gap in the program evaluation literature by describing how the form and function of rubrics assist program evaluators to reach explicitly evaluative conclusions.

Research Design

The three studies in the dissertation study explore the following:

Study 1) Rubrics in Program Evaluation

Research question: To what extent and how are rubric use and utility portrayed in the peer-reviewed literature on program evaluation?

Study 2) How Program Evaluators Use and Learn to Use Rubrics to Make Evaluative Reasoning Explicit

Research question: In what ways do evaluators use rubrics as evaluation-specific tools in program evaluation and how did they learn to do so?

Study 3) How the Form and Function of Rubrics Make Evaluative Reasoning Explicit

Research aim: To explicate, based on the logic and nature of evaluation, how the form and function of rubrics make evaluative reasoning explicit.

The methods for each of the stand-alone studies are specified separately in the following sections.

Study 1

The method for the first study is a literature review. The data for Study 1 are journal articles from peer-reviewed program evaluation journals. I selected published literature for two reasons: 1) The articles had been reviewed by evaluation peers signifying a standard of quality; and 2) Ease of access. To be retained for review, rubrics discussed in the articles had to be used in program evaluation activities, but could not be

used to measure the performance of individuals (e.g., students) or quality of products (e.g., curriculum). Twenty articles were retained for review.

Study 2

In the second study, interviews were conducted with authors of a) studies identified as relevant from the literature review for Study 1; b) authors of relevant professional development and instructional materials (e.g., evaluation books, workshops) identified in Study 2; and c) individuals identified by interviewees as having relevant experience (accomplished through a snowball sampling method). These individuals were interviewed to better understand how they use rubrics and how they learned to use rubrics.

Study 3

The third study concludes by explicating how the form and function of rubrics support evaluators in a core responsibility of their work, which is to draw conclusions that are transparent and explicitly evaluative.

Significance

Study 1 is the first systematic study of how rubrics are being portrayed in the program evaluation literature. It outlines how practitioners and scholars are using or discussing rubrics.

In Study 2, people who use rubrics in an evaluation-specific way were interviewed to better understand the nuances of how they use rubrics and how they

learned to use them. Findings from the study reveal gaps in training. Identifying the benefits and challenges of rubric use allows for deeper understanding of evaluation-specific rubric application and can possibly drive future focus in rubric related evaluator training.

In Study 3, a strong inferential approach is used, relying on the logic and nature of evaluation (Scriven, 1981), evaluator competency frameworks, as well as data presented in the earlier studies, as evidence that interpreting data to reach evaluative conclusions is a core function of the evaluator, and to advance understanding of how the form and function of rubrics can support evaluators in this enterprise.

Outline of the Dissertation

Chapter 1 introduces the project. It situates the preceding three chapters as stand-alone studies that together form a cohesive manuscript. Chapters 2 through 4 are stand-alone studies developed for submission as articles to peer-reviewed journals. Chapter 5 synthesizes the key findings from all studies and discusses the implications for evaluation theory, training, and practice.

References

Arens, S. A. (2005). *A study of evaluative reasoning in evaluative studies judged "outstanding."* An unpublished dissertation, Indiana University, Bloomington.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

Scriven, M. (1981). *Logic of evaluation*. Iverness, CA: Edgepress.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Value. (2016). In *Oxford dictionaries*. Editor © 2016 Oxford University Press. Retrieved from http://www.oxforddictionaries.com/definition/english/value

II. RUBRICS IN PROGRAM EVALUATION

Introduction

Evaluation is practiced in an array of programmatic contexts by evaluators hailing from diverse academic backgrounds, which distinguishes it from other professions with more consistent and clear-cut pathways into practice such as accountancy, law, or psychology. The fact that evaluators bring expertise from various disciplines affords great benefit if evaluators both leverage and share theory and practice from their respective fields. The American Evaluation Association (AEA; 2014) defines evaluation as "a systematic process to determine merit, worth, value or significance." Determining something's merit, worth, value, or significance is a complex task, which is often done implicitly, lacking transparent justification regarding reasoning (Arens, 2005) or left undone, providing evidence without drawing it together for evaluative conclusions (Davidson, 2013).

I have an education background, and in education we have a tool that we use regularly to evaluate student performance. This tool from education could also be used by program evaluators to make systematic evaluative judgments that are explicit by making their reasoning transparent. It is called a rubric. In this study, I review the extent to which and how rubrics, a fundamental tool used in education, are being discussed in program evaluation literature. I begin by exploring the background, or etymology, of the word rubric, first in a general sense and then within program evaluation, to determine the prevalence and evolution of the use of the term. I then present a review of literature to address the main research question:

To what extent and how are rubric use and utility portrayed in the peer-reviewed literature on program evaluation?

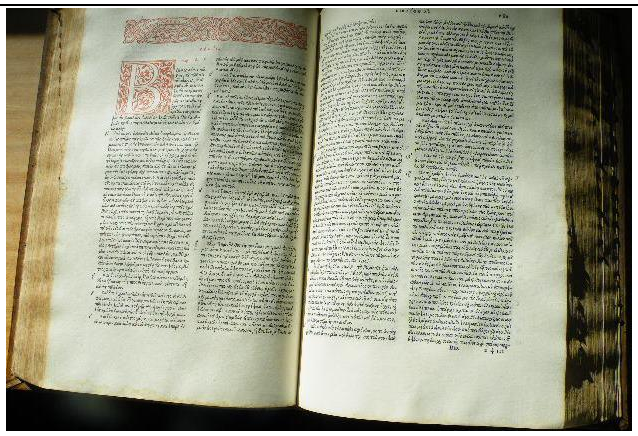
More specifically:

1) What is the trend in the use of the word in the program evaluation literature?
2) In what program evaluation contexts are rubric use and utility portrayed?
3) In what ways are rubrics being used or described as having utility in program evaluation?
4) How do evaluators develop, refine, and test rubrics?

Background

In keeping with Carl Sagan's (1980) statement, "You have to know the past to understand the present," I found myself tracing the term *rubric* back farther than I expected, but I continued because I found the results fascinating. The *Collins English Dictionary* (2013) states that in Latin the word for red is *ruber; rubrike* means red ochre and also came to signify red lettering; *rubrīca* means terra or red earth, and was the foundation for the term *ruddle,* which denotes the marking of a distinctive symbol on sheep using a smudge of red earth. *The Catholic Encyclopedia* (2009) dates the word *rubrīca* back to the "ancients" when carpenters used red earth to mark their cut, or saw line, on wood. Images of old, albeit less archaic, uses of the term rubric date back as far as the 9th century, when either a single large decorative red letter or the red lettering of a small sample of text was used to herald significant passages (Morgan & Panayotova, 2009). Popham (1997) noted that this literary practice (see Figure 1) originated from

monks who laboriously hand-copied manuscript pages, a practice that is also documented in the *Online Etymology Dictionary* (2014), which adds that, circa 1300, rubrics could further denote red scripted "directions in religious services."

a. 

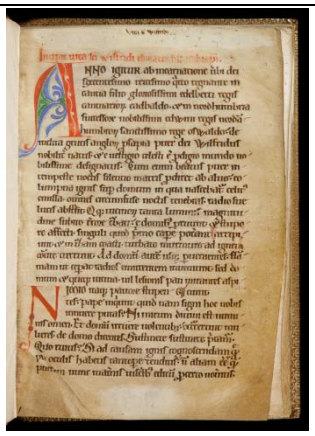
b. 

Figure 1. Two examples of early red typographical rubrics:

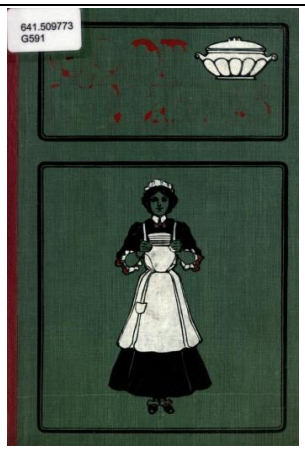
(a) Sacrae Scriptureae Veteris Novaeque Omnia. Photo by K. Martens (2015) with permission of and held by Dalton McCaughey Library, Ormond College, Melbourne, Victoria, Australia.

(b) Vita Sancti Wilfridi (Life of St Wilfrid) and Bede. Image from Historia ecclesiastica Folio 1r. England, the North, (2nd half of 12th Century). Held in Art Gallery of Ballarat's Crouch Manuscript collection. Digitized free access by the State Library of Victoria, Australia.

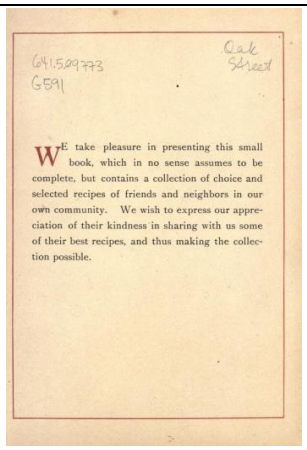
Secular manuscripts in the United States also used red typographical rubrics, as can be seen in the early 20th century example in Figure 2. The figure shows photographs of rubrics contained in the 1906 book *Good Recipes,* authored by The Woman's Society of the Winnetka Congregational Church. As was common for the time, the title of this publication appears in red lettering (notice the faded red title on the cover of the book in Figure 2, image a). The red lettering continues in the forward of the book, where the first

letter is quite large and printed in red typeface (Figure 2, image b). Figure 2, image c shows that the section headings of the book are also in red lettering. But it is the title, *Good Recipes*, that leads one to imagine that all recipes contained within had been carefully evaluated and selected as significant choices for publication—coincidence, or an early example of evaluation rubric use?

a.



b.



c.

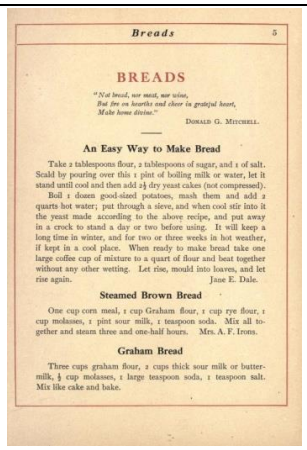


Figure 2. Three typographical rubric examples from a 1906 book with the evaluative title *Good Recipes:* (a) faded red title of the book on the cover; (b) large red W starting the paragraph; and (c) red section heading for breads. Reproduced from the freely available digital collection of University of Illinois Urbana-Champaign.

Use of rubrics as typographical elements waned in the mid-1900s, when traditional red lettering began to disappear as new, less expensive printing methods emerged such as using all black typeset that distinguishes significant text through bold or italic fonts. This change in practice is evidenced by William Faulkner, who noted in a 1929 letter to his agent that italics would get the point across but that he would prefer colored ink (Bleikasten, 1982, pp. 3–4).

A parallel evolution, beyond typeset, of the term rubric is in reference to classification spectrums. An example of use in this manner is the 1975 paper, *International Classification of Diseases-9*, published on the U.S. Centers for Disease Control and Prevention (2007) website. The paper outlines how Francois Bossier de Lacroix (1706 – 77) made the first systematic attempt at disease classification in the 18th century. The 1975 paper explains that in 1855, the second International Statistical Congress in Paris adopted a classification list that included "138 rubrics" that became the basis for the *Internal List of Causes of Death.* It is unclear if application of the word *rubric* to the 1855 work was a 1975 addition or if it had been used back in the 1800s in this manner. What is clear is that use of the word *rubric* as a classification term has become common among current day academics.

For a better understanding of this and other more recent uses of the word rubric and its prevalence of use across disciplines, I used the interdisciplinary database SCOPUS to search for appearances of the word rubric in article titles or abstracts from 1960 to 2014. Figure 3 shows the results, which demonstrate a substantial increase in use of the term in online articles (interpretation should be tempered by an expected increase due to the increase in the number of papers published online since 1999).

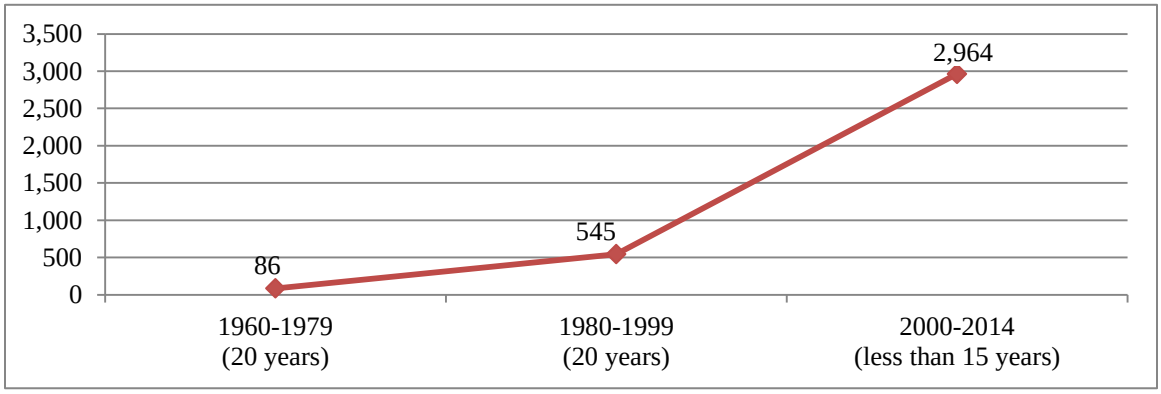
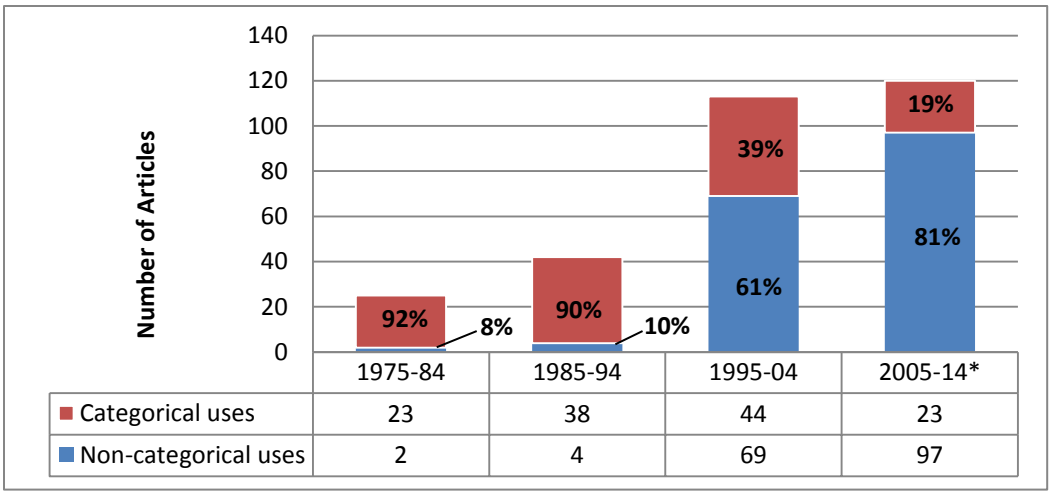


Figure 3. Number of article returns for search term *rubric* using the interdisciplinary database SCOPUS.

To find out if the use of the word *rubric* increased similarly in evaluation journals, I mirrored the search using 16 peer-reviewed evaluation journals with date parameters that reflected the earliest accessible online journal issue dates. As can be seen in Figure 4, the total frequency of use of the term rubric increased substantially in the decades since 1975—by almost fivefold. In conjunction, a review of all returns showed that the manner in which the term is used in evaluation has also changed dramatically. Between 1975 and 1994, the term *rubric* was used in the evaluation literature more than 90 percent of the time in the categorical or classification sense (e.g., "under the rubric of"). As overall use of the term increased, the categorical use of the term declined from 92 percent in 1975-84 to only 19 percent in 2005-14, although the absolute number of articles using the term in a categorical sense remained fairly stable, ranging from 23 articles in 1975-84 up to a high of 44 in 1995-2004 and back down to 23 in 2005-14. It is the increase in the non-categorical use of the term rubric that was of interest to me.



*2005-14 is just under 10 years, as it only spans through April 2014.

Figure 4. Number of evaluation articles[1] utilizing the term *rubric* by 10-year periods.

The increased use of the term in the literature in the 1990s is due in part to the influence of the increase in practice in education of using rubrics to evaluate student performance. This practice began in the 1970s. Early versions of education rubrics, called Standardized Developmental Ratings (SDRs) were used to assess developmental stages of learning through student writing and drawing samples (Dirlam & Byrne, 1978). SDRs evolved through a decade of work in the 1970s, but were originally based on Noam Chomky's (1957) review of Skinner's *Verbal Behavior,* in which Chomsky remarks on the lack of units of analysis for behavior (as cited in Dirlam, 2015). But it wasn't until the early '80s that the term rubric was applied to student assessment ratings, by Mel Grubb (1981) in his book, *Using Holistic Evaluation* (as cited in Edwards & Sailors, 2014).

Rubrics for evaluating performance are typically comprised of three basic components: 1) **criteria** of merit; 2) **standards** of merit; and 3) **descriptors** of

[1] Table 1 details 15 of the 16 evaluation journals accessed. Included in Figure 4, but excluded from Table 1 due to its policy focus is Educational Evaluation and Policy Analysis (EEPA).

performance (Popham, 1997; Reddy & Andrade, 2014). *Criteria* are defined for the purposes of this paper as non-overlapping dimensions or components of quality (Allen & Tanner, 2006; Scriven, 2000, revised 2005, 2007). *Standards* are the scaled levels of achievement or the spectrum of goodness on which the criteria will be assessed or rated (Allen & Tanner, 2006; Davidson, 2005; Dickinson & Adams, 2012). *Descriptors* provide examples of what quality would look like for each criterion at each performance level (Davidson, 2005). Figure 5 provides a visual depiction of how these three components can be organized in a matrix, although the configuration of the matrix can just as easily be inverted, with the standards down the far left column and the criteria across the top row (see Figure 9). Figure 6 shows an example of a rubric for marking essays (chosen because the standardization of essay marks was the origin of evaluation rubrics).

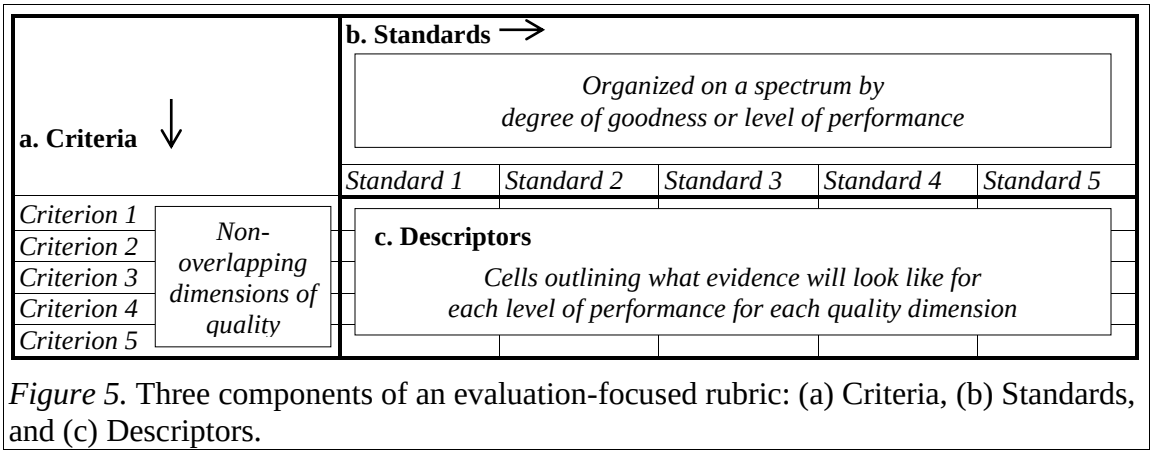


Figure 5. Three components of an evaluation-focused rubric: (a) Criteria, (b) Standards, and (c) Descriptors.

Criteria ↓	Standards →						
	6	5	4	3	2	1	0
Overall	Essay demonstrates clear and consistent mastery, although it may have a few minor errors	Essay shows reasonable consistent mastery, but has some errors or lapses in quality	Essay demonstrates adequate mastery, although it has lapses in quality	Essay demonstrates developing mastery marked by any or all of the below **weaknesses:**	Essay demonstrates little mastery and is flawed by any or all of the below **weaknesses:**	Essay demonstrates very little or no mastery and is severely flawed by below **weaknesses:**	Essay not as asked
Dimensions				(**weaknesses** refer to the below dimensional criterion rows)			
Reasoning	Develops effective insightful point of view; demonstrates outstanding critical thinking; uses clearly appropriate support evidence	Is well organized and focused, demonstrating coherence and progression of ideas	Develops a point of view demonstrating competent critical thinking, using adequate examples, reasons and other support evidence	Develops a point of view demonstrating some critical thinking, but may do so inconsistently or use inadequate support evidence	Develops a point of view that is vague; demonstrates weak critical thinking; provides inappropriate support evidence	Develops no viable point of view on the issue, or provides little or no evidence to support position	
Organization	Is well organized and clearly focused, demonstrating clear coherence and smooth progression	Is well organized and focused, demonstrating coherence and idea progression	Generally focused and organized, demonstrating some coherence and progression	Is limited in its organization or focus, or may demonstrate some lapses in coherence	Is poorly organized and/or focused, or demonstrates serious problems with coherence	Is disorganized or unfocused, resulting in a disjointed or incoherent essay	
Language	Exhibits skillful use of language, using a varied, accurate and apt vocabulary (vocab)	Exhibits facility in the use of language, using appropriate vocabulary	Exhibits adequate but inconsistent language; vocab is generally appropriate	Displays developing language use, but sometimes uses weak vocabulary	Displays very little facility in language, using very limited vocabulary or incorrect words	Displays fundamental errors in vocabulary	
Sentence structure	Sentence structure demonstrates meaningful variety	Sentence structure demonstrates variety	Sentence structure demonstrates some variety	Sentence structure demonstrates problems	Sentence structure demonstrates frequent problems	Demonstrates severe flaws in sentence structure	
Grammar, usage, mechanics	Is free of most errors	Is generally free of most errors	Has some errors	Contains an accumulation of errors	Contains errors so serious meaning is somewhat obscured	Contains pervasive errors so persistent meaning is masked	

Descriptors

Figure 6. Example rubric. Based on the SAT essay scoring guide "Essay scoring: How it's scored, and what the scores mean," by The College Board, 2015. Retrieved from: http://sat.collegeboard.org/scores/sat-essay-scoring-guide

In conjunction with the substantial uptick in references to rubrics in evaluation in the mid-90s was the release of Deborah Fournier's 1995 edited issue of *New Directions in Evaluation*, titled, *Reasoning in Evaluation: Inferential Links and Leaps*. This volume examined the "systematic means for arriving at evaluative conclusions, the principles that support inferences drawn by evaluators" (Fournier, 1995, p. 1). In this issue of *New Directions*, Michael Scriven (1995) argued that the most common deliverable of evaluations is evaluative claims. He went on to assert that evaluative claims are framed in the vocabulary of grading, ranking, scoring, or apportioning and concluded that the logic of evaluation requires explicit identification of criteria of merit (Scriven, 1995). In the same issue, Fournier said, "understanding the reasoning process to establish evaluative conclusions drawn in practice [is] the field's greatest unmet challenge" (Fournier, 1995, p. 1). An evaluation rubric is a tool that captures all the elements of the logic of evaluation as presented by Scriven and explicitly outlines the systematic reasoning called for by Fournier. To clarify, rubrics are developed to explicitly present the dimensions outlined by Scriven (evaluative claims and criteria of merit) and provide a systematic means for arriving at evaluative conclusions by applying the standards to the criteria of merit via performance definitions to justify the evaluative claims. Jane Davidson (2005) described rubrics as "a tool that provides an evaluative description of what performance or quality 'looks like' at each of two or more defined levels" (p. 247). Educators commonly use rubrics in this manner to assess student performance (Moskal, 2000; Popham, 1997), but less is known about how this practice has transferred to program evaluation.

In the remainder of this paper, I present the results of a literature review I conducted to better understand the extent to which and how rubrics are being used or described as having utility by evaluators in peer-reviewed program evaluation literature, especially to derive evaluative conclusions.

Method

This study focuses on how rubrics are presented in the peer-reviewed program evaluation literature. This means that other "semi-autonomous applied areas" as outlined by Michael Scriven (1991, p. 141; 1995, p. 50), such as evaluations of proposals, personnel, products, performance, and policies, are excluded. In this study, I selected published literature for reasons of quality (articles had been scrutinized and deemed meritorious by peers) and ease of access. The literature review I conducted focused on select program evaluation journals. My journal selection was initially informed by a list of evaluation journals compiled by Thomas Schwandt (2014). To ensure inclusion of the most relevant English-language program evaluation journals, I modified Schwandt's list of 15 evaluation journals by excluding one journal from the list, *Educational Evaluation and Policy Analysis*, because of its policy evaluation focus and added a journal to the list, *New Directions for Evaluation*. The final list included eight multidisciplinary evaluation journals, five education evaluation journals, and two health evaluation journals (see Table 1).

Table 1

Number of Articles including the Term Rubric by Evaluation Journal

	Journal	Start Year	Total Issues	Article Returns	Valid Articles*	Articles that Met Inclusion Criteria
1	*African Evaluation Journal*	2013	1	0	0	0
2	*American Journal of Evaluation*	1981	113	50	39	10
3	*Canadian Journal of Program Evaluation*	1986	71	1	1	0
4	*Evaluation and Program Planning*	1978	147	61	39	4
5	*Evaluation Journal of Australasia*	2001	24	2	2	2
6	*Evaluation Review: A Journal of Applied Social Research*	1977	215	19	19	0
7	*Evaluation: The International Journal of Theory, Research and Practice*	1995	75	18	17	0
8	*Journal of Multidisciplinary Evaluation*	2004	22	6	6	3
9	*New Directions in Evaluation*	1978	141	16	13	0
10	*Educational Research and Evaluation: An International Journal on Theory and Practice*	1995	103	38	37	0
11	*International Journal of Evaluation and Research in Education*	2012	8	0	0	0
12	*Practical Assessment, Research, and Evaluation*	1999	18	17	16	1
13	*Studies in Educational Evaluation*	1975	137	61	48	0
14	*Evaluation and the Health Professions*	1978	145	6	6	0
15	*Journal of Educational Evaluation for Health Professions*	2006	9	1	1	0
	Total article returns for *rubric*		**1,229**	**296**	**244**	**20**

*Number of valid original articles (excluding duplicate articles, book reviews, editorial notes, other non-article returns, articles in press, articles that were found not to contain the term, or that only contained the term in the references or solely in reference to use in other publications).

I searched all available online issues of the journals using the keyword *rubric*, which generated 296 article returns. I checked each article to determine its appropriateness for the study. My selection criteria included the following:

- The word *rubric* appeared in the body of the article, not just in references or footnotes.
- The word *rubric* was used in the evaluative sense; those that used the term strictly in the categorical manner (e.g., under the rubric of…) were excluded.
- Rubrics mentioned in the article were used for program or project evaluation; those that focused on policy, product (e.g., curriculum), personnel (e.g., principal or teacher), or performance (e.g., student) evaluation were excluded.
- The publication was a peer-reviewed article; book reviews, editorials, and other peripheral writings were excluded.
- The article was a unique contribution; those that referred to rubrics solely in reference to other published works were excluded.

In all, 20 articles were retained for analysis of the extent and nature of rubric portrayal in program evaluation. Figure 7 is a visual depiction of the filtering process.

Figure 7. Article filter flowchart.

Data from each article were recorded in an Access database. Areas of interest were frequency of use of the word rubric; context; how rubrics were used or described as having utility in program evaluation; and how rubrics were developed, refined, and tested. I conducted a narrative analysis to provide both a historical account of use and to identify the interconnections of described rubric use and utility in program evaluation.

A limitation of the study is that rubric use may be more widespread in practice than suggested by the published literature. Perhaps rubric use is just not a topic that

makes it into the peer-reviewed evaluation literature. Therefore, the results may not be an accurate reflection of what occurs in evaluation practice.

Results

The results of my analysis are organized by research sub-questions. Of note, the variables of interest were not uniformly present in the 20 articles.

Prevalence

The articles varied in the degree to which the authors focused on rubrics. A rough indicator of how important rubrics were to each publication is the number of times the word *rubric* appeared in the article. The frequency of use ranged from *1 time* (Brandon, Smith, Ofir, & Noordeloos, 2014; Braverman, 2013; Petersen, 2002; Roberts-Gray, Gingiss, & Boerm, 2007;) to *133 times* (King, McKegg, Oakden, & Wehipeihana, 2013) in a single article. Figure 8 shows a substantial increase since 2004 in both the number of times the term *rubric* was used in individual articles (blue line) and the dispersion of the 20 articles contained in this study (red line).

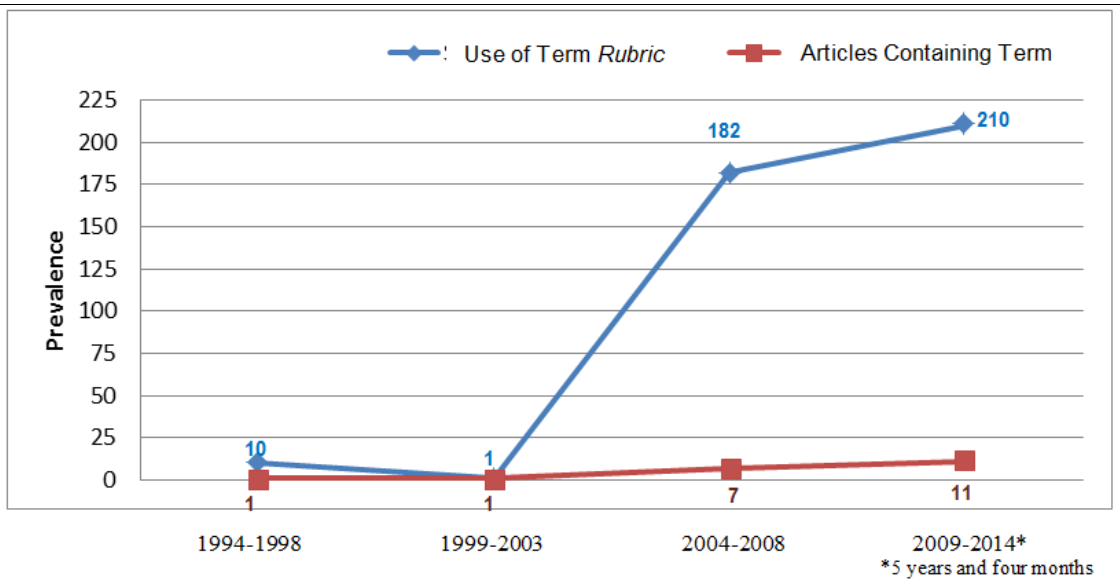


Figure 8. Frequency of use of the term *rubric* in articles over 5-year periods.

Context

The articles included in the review discussed rubric use predominately in education and health program evaluation, although the programs were quite diverse. The program types described in the articles included evaluation capacity building initiatives, multisite and multi-intervention physical activity and nutrition health promotion, planning and managing preparedness for emergencies and disasters, university improvement planning, professional development for African female scientists in agriculture, K-12 school improvement interventions, organizational capacity building in nonprofits, substance abuse prevention, system-wide education reform, tobacco prevention, and food security in international development. In 9 of the 20 articles, the authors included displays (7 articles) or abridged displays (2 articles) of the rubrics used in their work.

Rubric Use or Described Utility

The articles' authors described using rubrics for a variety of reasons, such as transforming data from one form to another (e.g., Davey, Gugiu & Coryn, 2010), to characterize organizational functioning (e.g., Gadja & Koliba, 2007), and to derive explicitly evaluative conclusions from the synthesis of evidence and values (e.g., Dickinson & Adams, 2012). Values used in evaluation are drawn from sources such as a needs assessment, literature, professional standards, or a logical analysis of the function of the evaluand (Scriven, 1991), they are not individually held values. Authors also described rubrics being used in various phases of inquiry such as data collection (7 articles), data analysis (7 articles), and synthesizing findings into conclusions (10 articles). Examples of types of use at varying stages of inquiry were to record performance judgments during site visits (data collection), to code data (data analysis), and to synthesize evidence into evaluative conclusions during the evaluation interpretation or synthesis phase (some conclusions were not evaluative). Some authors used rubrics in more than one way and in more than one phase of their studies. For example, Clements (2012) used a rubric to transform observational and interview data into numerical codes during data collection, then to analyze the data to show the level of change between reconstructed baseline and current functioning, and to synthesize the data into evaluative claims.[2]

I found that rubrics in the early study phases were more often used to transform data from one form to another, such as, to transform observations in the field (visual data)

[2] The uses at different stages are not explicitly outlined in the article, but as part of the evaluation team working with Clements (and noted as such in the article), I have intimate knowledge of the way we used rubrics in this study.

into written codes (e.g., Brandon, Smith, Ofir, & Noordeloos, 2014; Clements, 2012). Predominately, rubrics in the reviewed articles were used in the data collection and analysis phases to transform qualitative data (e.g., from interviews, observations, documents, open-ended items in questionnaires) into codes (most often scores) to allow for quantitative analysis (e.g., Roberts-Gray, Gingiss & Boerm, 2007). For example, Petersen (2002) scored key informant interview responses using a rubric prior to aggregation. In later phases of studies, rubrics were sometimes used to synthesize data collected via multiple methods (e.g., document review, survey, interview, observational data; Clinton, 2014) or to synthesize quantitative and qualitative data. Figure 9 shows an example of synthesis of quantitative and qualitative data into qualitative judgments using evaluative terms (i.e., excellent, very good, good, poor).

TABLE 1: MERIT DETERMINATION FOR QUALITY—THREE-DAY AND ONE-DAY WORKSHOPS

Excellent	Nearly all participants (95%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Nearly all would recommend the workshops to others and no problems were identified regarding the quality of workshop content, design or delivery.
Very Good	Most participants (75–94%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Most would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Good	At least half of participants (50–74%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. At least half would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Poor	Less than half (<50%) of participants either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Less than half would recommend the workshops to others OR major problems were identified regarding the quality of workshop content, design or delivery and these were not managed effectively.

Figure 9. Example rubric of synthesis of methods data into qualitative judgments from "Building Evaluation Capability in the Public Health Workforce: Are Evaluation Training Workshops Effective and What Else Is Needed?" by P. Dickinson and J. Adams, 2012, *Evaluation Journal of Australasia, 12*(2), p. 32. Reprinted with permission.

Rubric Characteristics

In this section, I describe the characteristics of the rubrics in the review, as well as the processes described by the authors for developing, refining, and testing them. The characteristics are organized by the three elements of an evaluation rubric (criteria, standards, and descriptors). The descriptions of rubric characteristics varied in level of detail across the articles.

Criteria. All authors described criteria of *merit* (quality) as opposed to worth (e.g., cost-effectiveness or some sort of value for money) or significance (importance). The authors of six articles provided overarching definitions or examples of the criteria expressed in their rubrics. For example, Clements (2012) described a rubric[3] for scoring nutritional gains that contained, among other criteria, a criterion for nutritional staples. Nutritional staples were defined as including "cereals (wheat, barley, rye, maize, or rice), starchy root vegetables (potatoes, yams, taro, and cassava), and special fruits (breadfruit and plantains)" (Clements, 2012, p. 23). In this example, the presence of nutritional staples is a marker of a quality food security program.

All authors, but one, described or presented *analytic* rubrics, meaning the rubrics were developed to evaluate performance on more than one criterion of merit (similar to the earlier example in Figure 6). Alternatively, Figure 10 shows the one *holistic* rubric, different because it had one overarching criterion of merit (i.e., quality), instead of two or more components or dimensions of merit.

[3] Clements attributes the development of the instrument to Dr. Michael Scriven, the principal investigator of the study, and Scriven's evaluation team.

Descriptors. Descriptors are the interior cells of a rubric that outline what performance looks like across the standard spectrum by criterion (refer back to Figures 5 and 6). Descriptors may be stated quantitatively, qualitatively, or in a mixture of both ways. Continuing with the example from Clements (2012, p. 23), the descriptors for the *nutritional staples criterion* were a mixture of quantitative and qualitative performance examples as shown in Figure 10. Also of note, Figure 10 is an example of standards (ratings 0-5) organized down the far left column and the criterion (nutritional staples) across the top row. This is counter to the presentations in Figure 5 and 6 but matches the presentation in Figure 9. I chose this configuration for Figure 10 simply because it is the best use of space on a portrait oriented letter (or A4) sized sheet of paper.

	Nutritional Staples: (wheat, barley, rye, maize, or rice), starchy root vegetables (potatoes, yams, taro, and cassava), and special fruits (breadfruit and plantains)
5	Enough staples year-round
4	Three meals of staples per day year-round, but occasionally not enough to satisfy
3	Two or three meals of staples per day year-round, but at least once a week, or for some period, such as a hungry season, not enough to satisfy
2	Mild shortage of staples year-round
1	One meal of staples per day through much of the year, or two meals but usually not enough to satisfy
0	No access to staples

Figure 10. Example of quantitative and qualitative descriptors. Based on textual information found in "Evaluating the cost effectiveness of heifer international country programs," by P. Clements, 2012, *Journal of Multidisciplinary Evaluation, 8*(18), p. 23.

All rubrics in the reviewed articles but one were developed for a specific evaluation context, whether that was a specific program or project (e.g., Heifer Projects International) or a particular type of program (e.g., partnering). These rubrics include

context-specific detail in the interior cells of the rubric that limits their use outside of the setting for which they were developed. In the above example from the Clements article, the descriptors (interior cells containing examples of what performance would look like for each criterion at the different levels of performance) were developed for a specific international development program to judge families' food security, but the rubric was to be used in multiple countries over a span of many years. So the rubric descriptors were developed to be specialized for this specific program, which included many projects around the world—a tricky balance. The other rubrics found in the review were similarly developed to be specific to their evaluation context, except one that was developed with generic descriptors. Generic descriptors enable users to apply the rubric to just about any program or context. Figure 11 shows the one generic rubric contained in the review, which is also the one holistic rubric described in the review. Finally, the descriptors were all phrased in absolute terms. This type of determination states quality independent of other performances (as opposed to relative determinations that outline quality compared to like performances). This means an absolute *good* is good no matter what, whereas a relative *good* is good in comparison to others.

Table 1
Example of a Generic Rubric

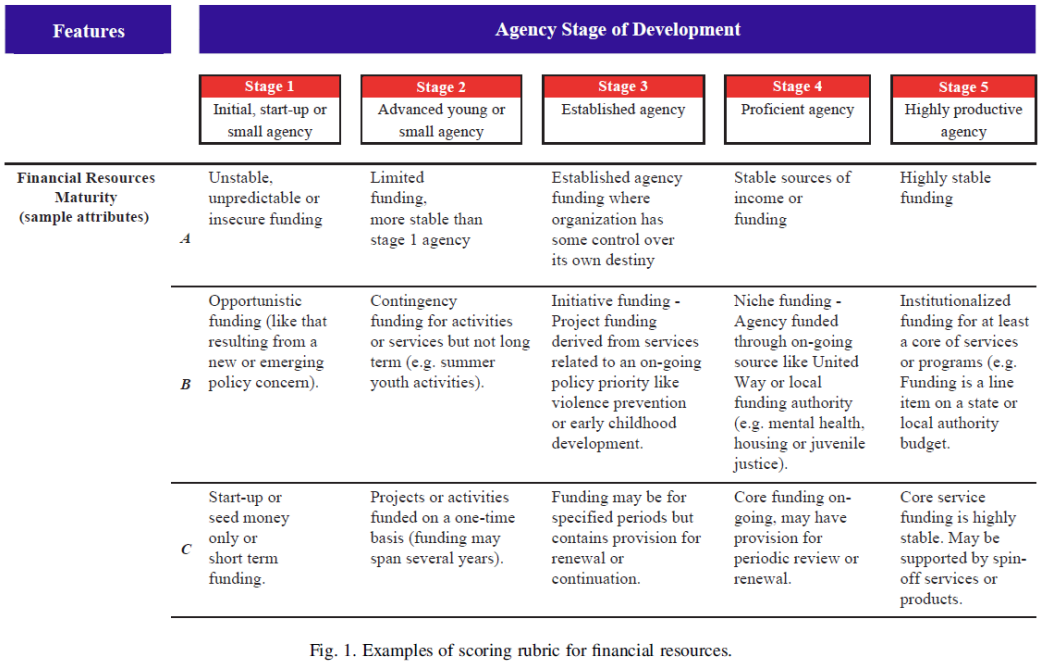
	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

Figure 11. Example holistic generic rubric from "Evaluative Rubrics: A Method for Surfacing Values and Improving the Credibility of Evaluation," by J. King, K. McKegg, J. Oakden, and N. Wehipeihana, 2013, *Journal of Multidisciplinary Evaluation, 9*(21), p. 14. Reprinted with permission.

Standards. In conjunction with the descriptors in all articles being presented or described in absolute terms, so are the standards (as opposed to relative standards that outline a direct comparison, such as superior, average, inferior; Davidson, 2005).The absolute standards contained in the review were presented as scores (e.g., 0–4 points), grades (e.g., A, B, C, D, E), evaluative terms (e.g., excellent, good, poor), and levels of met criteria (e.g., fully met, met, not met). Of note, authors working in organizational development evaluation (e.g., with nonprofits or inter- or intra-agency collaborations)

presented absolute scoring rubrics (using ordinal numbers) that also included developmental terms to describe performance, such as *levels of integration*, *stages of development*, or *levels of collaboration* (e.g., networking, cooperating, partnering, unifying). One such set of authors noted that their developmental ratings were not meant to be evaluative judgments, but instead the rubric was a "taxonomy of observed characteristics or symptoms" (Schuh & Leviton, 2006, p. 174), but because the rubric displayed and described in the article (Figure 12) contained all the elements of an evaluation rubric (standards, criteria, and descriptors), the article was retained in the review.



Features		Agency Stage of Development				
		Stage 1 Initial, start-up or small agency	Stage 2 Advanced young or small agency	Stage 3 Established agency	Stage 4 Proficient agency	Stage 5 Highly productive agency
Financial Resources Maturity (sample attributes)	*A*	Unstable, unpredictable or insecure funding	Limited funding, more stable than stage 1 agency	Established agency funding where organization has some control over its own destiny	Stable sources of income or funding	Highly stable funding
	B	Opportunistic funding (like that resulting from a new or emerging policy concern).	Contingency funding for activities or services but not long term (e.g. summer youth activities).	Initiative funding - Project funding derived from services related to an on-going policy priority like violence prevention or early childhood development.	Niche funding - Agency funded through on-going source like United Way or local funding authority (e.g. mental health, housing or juvenile justice).	Institutionalized funding for at least a core of services or programs (e.g. Funding is a line item on a state or local authority budget.
	C	Start-up or seed money only or short term funding.	Projects or activities funded on a one-time basis (funding may span several years).	Funding may be for specified periods but contains provision for renewal or continuation.	Core funding on-going, may have provision for periodic review or renewal.	Core service funding is highly stable. May be supported by spin-off services or products.

Fig. 1. Examples of scoring rubric for financial resources.

Figure 12. Example developmental rubric. The criterion presented is under the criterion, Financial Resources. Other criteria outlined in the article, but not shown in the rubric, are Governance, Organizational Development, Internal Operations, and Core Services. From "A Framework to Assess the Development and Capacity of Non-Profit Agencies," by R. G. Schuh and L. C. Leviton, 2006, *Evaluation and Program Planning*, *29*(2), p. 175. Reprinted with permission.

Table 2 shows the characteristics of rubrics included in the review.

Table 2

Summary Characteristics of Rubrics in Reviewed Articles

Characteristic	Number of articles (N=20)
Criteria	
Analytic	15
Holistic	1
Generic	1
Specific	15
Standards	
Absolute* ∞	19
Numeric (e.g., scoring; numerically named stages)	15
Developmental (e.g., stages)	7
Evaluative terms (e.g., excellent, good, adequate, poor)	4
Level terms (high, partial, low; met; partially met; unmet)	2
Grading (e.g., A, B, C, D, E)	1
Relative	0
Descriptors	
Generic	1
Specific	13
Absolute	14
Relative	0

∗ Some articles were marked in more than one category (e.g., numeric and developmental)
∞ If no other information was provided, a scoring rubric is deemed absolute

Process. In 13 of the 20 articles, authors described the rubric development process. The process occurred in two ways: 1) as an activity limited to the evaluation team's input based on research (e.g., needs assessment findings; analysis of preliminary data), theory, and/or relevant literature; and 2) as a participatory activity that ranged from stakeholder input on the refinement of an evaluator-developed rubric to fully engaging stakeholders in the process from start to finish.

Validity and reliability. In eight articles, the authors noted the importance of calibrating raters or outlined how to calibrate raters to increase interrater reliability. For example, Stemler (2004) based his entire article on interrater reliability estimates of multiple raters who use rubrics to make independent judgments of performance. He compares consensus, consistency and measurement approaches. But for most articles, the authors only mention that the judges were provided trainings (3 articles) and/or refinements were made in the field (4 articles). Rubrics were mentioned as being pilot tested in five articles.

Discussion

The definition of rubrics set out in this study (containing criteria, standards, and descriptors) makes them inherently evaluative because judgments of quality are made by combining evidence and value by plotting actual performance onto example descriptions of performance where each criterion of merit meets each standard of merit. So, if by definition the rubrics are evaluative, it does not matter the type of evaluative claim made, how the rubrics were developed, or the phase of the study in which they were employed; they are by design going to allow the user to make explicit judgments. Yet, half of the authors described them as common data collection and measurement tools, tasks that require classification (such as coding), that are not evaluative. While rubrics are useful for these purposes, these uses fall short of harnessing the power of rubrics for evaluation. In half of the articles, authors clearly described using rubrics to reach evaluation conclusions, with four of the ten using explicitly evaluative terms in the judgments, and an additional two using scores to denote merit—in other words, six articles purposefully

used rubrics as an *evaluation-specific methodology*. I find this number, six articles, troubling because, as noted in the opening paragraph of this review, evaluation is the systematic determination of merit, worth, value, or significance[4] (AEA, 2014). This review of program evaluation literature spanned close to 40 years and initially included 239 articles, resulting in only 20 examples of rubric use or described utility in program evaluation, with a paltry 6 articles describing rubric use as an evaluation-specific methodology. This review suggests that the use of rubrics in program evaluation is somewhat rare, although, mentioned as a limitation is possibility that rubric use in program evaluation is actually more prevalent in practice, but is just not a topic that makes it into the peer reviewed literature. If this is the case, and it is due to a schism between the practical (non-academic) nature of rubrics and the academic (peer-reviewed) perspective, the issue highlights the importance of conducting research on evaluation practice. Alternatively, if evaluation practitioners are missing the opportunity to share their experience and guidance on an evaluation-specific methodology, specifically, the art and science of rubric development and use, a perspective that would be welcomed in scholarly journals, then that in itself is a finding worth noting.

[4] While some evaluators make the case that it is not the job of the evaluator to make these types of value claims, not doing so is at odds with AEA's definition of evaluation.

References

References marked with an asterisk indicate studies included in the analysis.

*Adams, J., & Dickinson, P. (2010). Evaluation training to build capability in the community and public health workforce. *American Journal of Evaluation, 31*(3), 421-433.

Allen, D., & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *Cell Biology Education—Life Sciences Education,* 5, 197-203.

American Evaluation Association. (2014, January 10). *What is Evaluation?* Retrieved from http://www.eval.org/p/bl/et/blogid=2&blogaid=4

Arens, S. A. (2005). *A study of evaluative reasoning in evaluative studies judged "outstanding."* An unpublished dissertation, Indiana University, Bloomington.

Bleikasten, A. (1982). *William Faulkner's the sound and the fury: A critical casebook.* New York: Garland.

*Brandon, P. R., Smith, N. L., Ofir, Z., & Noordeloos, M. (2014). Monitoring and evaluation of African women in agricultural research and development (AWARD): An exemplar of managing for impact in development evaluation. *American Journal of Evaluation, 35*(1), 128-143.

*Braverman, M. T. (2013). Negotiating measurement: Methodological and interpersonal considerations in the choice and interpretation of instruments. *American Journal of Evaluation, 34*(1), 99-114.

*Clements, P. (2012). Evaluating the cost effectiveness of Heifer International country programs. *Journal of Multidisciplinary Evaluation, 8*(18), 16-28.

*Clinton, J. (2014). The true impact of evaluation: Motivation for ECB. *American Journal of Evaluation, 35*(1), 120-127.

The College Board. (2015). Essay scoring: How it's scored, and what the scores mean. Retrieved from: http://sat.collegeboard.org/scores/sat-essay-scoring-guide

*Coryn, C. L. S., Gugiu, P. C., Davidson, E. J., & Schröter, D. C. (2007). Needs assessment in hidden populations using respondent-driven sampling. *Evaluation Journal of Australasia, 7*(2), 3-11.

*Davey, J. W., Gugiu, P. C., & Coryn, C. L. S. (2010). Quantitative methods for estimating the reliability of qualitative data. *Journal of Multidisciplinary Evaluation, 6*(13), 140-162.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davidson, E. J. (2013). *Actionable evaluation basics: Getting succinct answers to the most important questions.* Auckland, New Zealand: Real Evaluation.

*Dickinson, P., & Adams, J. (2012). Building evaluation capability in the public health workforce: Are evaluation training workshops effective and what else is needed? *Evaluation Journal of Australasia, 12*(2), 28-39.

Dirlam, D., & Byrne, M. (1978, March). *Standardized Developmental Ratings*. Paper presented at the annual 62nd meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Dirlam, D. (2015). *Changing wisdoms: History of developmental rubrics.* Ed. David K. Dirlam. Retrieved from: http://changingwisdoms.org/?s=Developmental%20Rubrics&sid=2&ss=History&ssid=1

Edwards, V., & Sailors, K. (2014). Creating new media rubrics: Quality student products for the 21st century. In H. H. Jacobs (Ed.), *Mastering Media Literacy* (pp. 49-66). Bloomington, IN: Solution Tree Press.

Fournier, D. (1995). Editor's notes. *New Directions for Evaluation, 1995*(68), 1-4.

*Gajda, R. (2004). Utilizing collaboration theory to evaluate strategic alliances. *American Journal of Evaluation, 25*(1), 65-77.

*Gajda, R., & Koliba, C. (2007). Evaluating the imperative of intraorganizational collaboration: A school improvement perspective. *American Journal of Evaluation, 28*(1), 26-44.

*King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), 11-20.

Martens, K. (Photographer). (2015). *Sacrae scriptureae veteris novaeque omnia* [photograph]. Melbourne, Victoria, Australia: Dalton McCaughey Library, Ormond College.

Morgan, N., & Panayotova, S. (2009). *A catalogue of Western book illumination in the Fitzwilliam Museum and the Cambridge Colleges, part I: The Frankish Kingdoms, Northern Netherlands, Germany, Bohemia, Hungary, Austria, The Meuse Region and Southern Netherlands*. London and Turnhout: Harvey Miller/Brepols.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research and Evaluation, 7*(3), n.p.

*Petersen, D. M. (2002). The potential of social capital measures in the evaluation of comprehensive community-based health initiatives. *American Journal of Evaluation, 23*(1), 55-64.

*Piper, D., Stein-Seroussi, A., Flewelling, R., Orwin, R. G., & Buchanan, R. (2012). Assessing state substance abuse prevention infrastructure through the lens of CSAP's Strategic Prevention Framework. *Evaluation and Program Planning, 35*(1), 66-77.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership, 55*(2), 72-75.

Reddy, Y. M., & Andrade, H. (2014). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.

*Roberts-Gray, C., Gingiss, P. M., & Boerm, M. (2007). Evaluating school capacity to implement new programs. *Evaluation and Program Planning, 30*(3), 247-257.

Rubric. (2009). *Catholic encyclopedia*. Editor © 2009 Kevin Knight. Retrieved from http://www.newadvent.org/cathen/13216a.htm

Rubric. (2013). In *Collins English dictionary*. Retrieved from http://www.collinsdictionary.com/dictionary/english/rubric

Rubric. (2014). In *Online etymology dictionary*. Editor © 2001-2014 Douglas Harper. Retrieved from http://www.etymonline.com/index.php?term=rubric

Sagan, C. E. (author and presenter). (1980). Episode 2: One Voice in the Cosmic Fugue [Television series episode]. In Adrian Malone (Producer), Cosmos: A Personal Voyage. Arlington, VA: Public Broadcasting Service.

*Schuh, R. G., Eichelberger, R. T., Stebbins, S., Pomer, B., Duran, L., Mahoney, J. F., Keane, C., Lin, C. J., & Potter, M. A. (2012). Developing a measure of local agency adaptation to emergencies: A metric. *Evaluation and Program Planning, 35*(4), 473-480.

*Schuh, R. G., & Leviton, L. C. (2006). A framework to assess the development and capacity of non-profit agencies. *Evaluation and Program Planning, 29*(2), 171-179.

Schwandt, T. A. (2014, March 25). Webinar in CBD special series: Session 1, What journals want [Digital recording]. AEA coffee break series. Retrieved from http://comm.eval.org/coffee_break_webinars/resources/viewdocument/?DocumentKey=bc8b3486-9a72-43ce-b604-7f7e6b5c224f

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

*Scriven, M. (1994). The final synthesis. *American Journal of Evaluation, 15*(3), 367-382.

Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions for Evaluation, 1995*(68), 49-70.

Scriven, M. (2000, revised 2005, 2007). The logic and methodology of checklists. Western Michigan University. Retrieved from https://www.wmich.edu/sites/default/files/attachments/u350/2014/logic%26methodology_dec07.pdf

*Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*(4), n.p.

*Supovitz, J. A., & Taylor, B. S. (2005). Systemic education evaluation: Evaluating the impact of system-wide reform in education. *American Journal of Evaluation, 26*(2), 204-230.

U. S. Centers for Disease Control and Prevention. (2007). *International Classification of Diseases – 9 (1975)*. Retrieved from http://wonder.cdc.gov/wonder/sci_data/codes/icd9/type_txt/icd9.asp

Vita Sancti Wilfridi (Life of St Wilfrid) and Bede. Excerpts from Historia ecclesiastica Folio 1r. England, the North, (2nd half of 12th Century). Held in Art Gallery of Ballarat's Crouch Manuscript collection. Digitized free access by the State Library of Victoria, Australia. Retrieved from http://search.slv.vic.gov.au/primo_library/libweb/action/search.do;jsessionid=3EBE14BF2239FB601EA30375ECEFD6EF?fn=search&ct=search&initialSearch=true&mode=Basic&tab=default_tab&indx=1&dum=true&srt=rank&vid=MAIN&frbg=&vl%28freeText0%29=Vita+Sancti+Wilfridi+%28Life+of+St+Wilfrid%29+and+Bede.+Excerpts+from+Historia+ecclesiastica+Folio+1r.+England%2C+the+North&scp.scps=scope%3A%28SLV_VOYAGER%29%2Cscope%3A%28SLV_DIGITOOL%29%2Cscope%3A%28SLVPRIMO%29&vl%281UIStartWith0%29=contains&vl%2810247183UI0%29=any&vl%2810247183UI0%29=title&vl%2810247183UI0%29=any

*Woodland, R. H., & Hutton, M. S. (2012). Evaluating organizational collaborations: Suggested entry points and strategies. *American Journal of Evaluation*, *33*(3), 366-383.

The Woman's Society of the Winnetka Congregational Church. (1906). *Good recipes*. Winnetka, IL. Digitized and held by University of Chicago, Urbana-Champaign.

Retrieved from http://www.archive.org/stream/goodrecipes00winn#page/n7/mode/2up

III. HOW PROGRAM EVALUATORS USE AND LEARN TO USE RUBRICS TO MAKE EVALUATIVE REASONING EXPLICIT

Introduction

Research is the systematic gathering of evidence for the advancement of knowledge (Merriam-Webster, 2015). Evaluation also involves systematically gathering evidence, but the evidence is used to determine merit, worth, or significance (American Evaluation Association, 2014). If both research and evaluation require the systematic collection of evidence, it is the reasoning of combining evidence *with values* that enables evaluators to determine merit, worth, or significance (Davidson, 2005). Relevant values in evaluation can be identified from sources such as a needs assessment, relevant literature, professional standards, or a logical analysis of the function of the evaluand—they are not individually held values (Scriven, 1991). This process is one that is truly unique to evaluation (Scriven, 1991), but unfortunately, little guidance is provided about how to combine evidence and values to reach explicitly evaluative conclusion in a transparent and credible way (Davidson, 2005). A rubric is a tool that educators have been using to combine evidence with values to assess student performance since the 1970s (Dirlam, 2015). Program evaluators can use the same tool, in the same way, to transparently draw explicitly evaluative conclusions. However, examples of the use of rubrics in this manner are rare in the evaluation literature (Martens, 2015).

A rubric for combining evidence and values is defined for the purposes of this study as containing three components: 1) *criteria* of merit—the aspects of an evaluand that will be examined to determine its quality; 2) *standards* or a spectrum of goodness—how the thing being evaluated will be labeled in levels of quality (e.g., grades, scores,

value statements); and 3) *descriptors* or examples of what performance looks like for each criterion at each performance level (Popham, 1997; Reddy & Andrade, 2014). Figure 13 shows an example of how the three components are commonly combined.

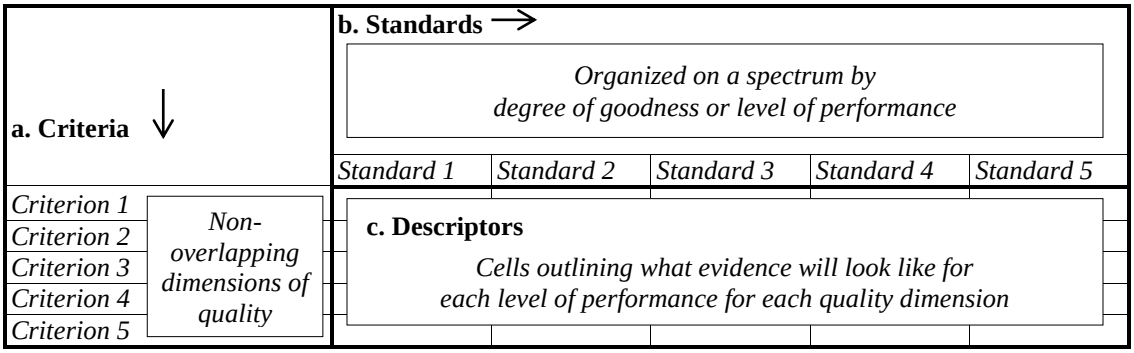


Figure 13. Three components of an evaluation-focused rubric: (a) Criteria, (b) Standards, and (c) Descriptors. From "Rubrics in Program Evaluation," by K. S. R. Martens, 2015, unpublished manuscript, p. 14. Reprinted with permission.

In an earlier study, I found that when rubrics are used in program evaluation, it is commonly in one of two ways: 1) to transform data, a common non-evaluative use to, for example, classify or categorize; or 2) as an evaluation-specific methodology (Martens, 2015). As a common research methodology, rubrics are used in a non-evaluative manner, such as to present a set of rules to guide translation from one type of information into another. Examples of rubric data transformations typical to research are transforming visual data (observations) to codes during data collection, transforming qualitative data to scores for data analysis, and synthesizing quantitative and qualitative data together as occurs in some systematic reviews of literature. When rubrics are used as an evaluation-specific methodology (a term coined by Michael Scriven, 1991), they are used to blend "descriptive data with relevant values to draw explicitly evaluative conclusions"

(Davidson, 2005, p. xii). While the use of rubrics as a research tool is important, it is the evaluation-specific methodology of rubric use in program evaluation that I investigate in this study. This study is guided by the following research questions:

1) In what ways do evaluators use rubrics as evaluation-specific tools in program evaluation?
2) How did they learn to do so?

Method

I designed a naturalistic phenomenological study to investigate the lived experiences of the study participants to better understand the essence of the shared rubric experience in program evaluation (Creswell, 2014). To this end, I conducted key informant interviews. Key informant interviews are in-depth semi-structured conversations with a limited number of individuals (often 15-35 people) who have specialized experience in an area of interest (USAID, 1996), which in this study was experience in the development and use of evaluation-specific rubrics in program evaluation. The interviews were conducted with individuals identified in two ways: 1) through a previous literature review I conducted of peer-reviewed program evaluation journals (Martens, 2015), which produced few examples of relevant rubric use, and 2) my examination of instructional materials for this study, to purposefully include practitioners who may be working with rubrics, but are not publishing in the peer-reviewed literature.

I identified potential informants who published instructional materials by examining 20 top-selling evaluation books from five publishing houses (see Table 3) and by examining evaluator professional development materials (e.g., conference presentation abstracts) housed on four professional evaluation association and society websites.

Table 3.

Twenty Top-Selling Evaluation Books in Alphabetical Order

Author	Year	Title
Alkin, M. C.	2011	*Evaluation essentials: From A to Z*
Bamberger, M., Rugh, J., & Mabry, L.	2011	*Real world evaluation: Working under budget, time, data, and political constraints* (2nd ed.)
Chen, H. T.	2014	*Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective* (2nd ed.)
*Davidson, E. J.	2005	*Evaluation methodology basics: The nuts and bolts of sound evaluation*
Fetterman, D. M.	2001	*Foundations of empowerment evaluation*
∞Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R.	2010	*Program evaluation: Alternative approaches and practical guidelines* (4th ed.)
McDavid, J. C., Huse, I, & Hawthorn, L. R. L.	2013	*Program evaluation and performance measurement: An introduction to practice* (2nd ed.)
*Mertens, D. M. & Wilson, A. T.	2012	*Program evaluation theory and practice: A comprehensive guide*
*Owen, J. M.	2006	*Program evaluation: Forms and approaches* (3rd ed.)
Patton, M. Q.	2008	*Utilization-focused evaluation* (4th ed.)
Patton, M. Q.	2010	*Developmental evaluation: Applying complexity concepts to enhance innovation and use* (1st ed.)
Patton, M. Q.	2011	*Essentials of utilization-focused evaluation*
Posovac, E. J., & Carey, R. G.	2011	*Program evaluation: Methods and case studies* (8th ed.)
Rossi, P. H., Lipsey, M. W., & Freeman, H. E.	2004	*Evaluation: A systematic approach* (7th ed.)
Royse, D., Thyer, B. A., Padgett, D. K.	2016	*Program evaluation: An introduction to an evidence-based approach.* (6th ed.)
Scriven, M.	1991	*Evaluation thesaurus* (4th ed.)
*Stufflebeam, D. L., & Coryn, C. L. S.	2014	*Evaluation theory, models, and applications* (2nd ed.)
Weiss, C. H.	1997	*Evaluation: Methods for studying programs and policies* (2nd ed.)
*Wholey, J. S., Hatry, H. P., & Newcomer, K. E.	2010	*Handbook of practical program evaluation* (3rd ed.)
Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F.	2011	*The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.)

*Also available as an e-book

∞Not available for an Amazon.com "Look Inside" or Google books electronic search

I chose evaluation books because the essence of evaluation is merit determination, so it follows that this topic should be an important topic covered in evaluation books,

including tools that explicitly and transparently assist evaluators to judge merit (e.g., rubrics). I compiled the list of books by contacting three publishing houses. The lists I received overlapped substantially and were consistent with a list I compiled from a Scopus database search.

Using the search term *rubric*, I was able to search all books electronically except one, which I reviewed manually by using the book's index. All returns for the search term were reviewed to determine if the authors discussed rubrics as an evaluation-specific methodology (excluding authors who solely referred to rubric use found in the work of other authors' publications).

I also identified potential informants by examining evaluator professional development materials (e.g., conference presentation abstracts) housed on four professional evaluation organization websites. These four organizations were chosen because they have the largest English language membership base. I used the search engines contained in the websites of the following professional evaluation organizations:

- American Evaluation Association (AEA)
- Australasian Evaluation Society (AES)
- Canadian Evaluation Society (CES)
- European Evaluation Society (EES)

In total, the peer-reviewed literature from a previous study (Martens, 2015) identified eight individuals who discussed the use of rubrics as an evaluation-specific tool. The instructional materials (books and professional association materials) identified three additional individuals (who had not already been identified in the literature review) for a total of eleven potential key informants.

Interview Participants

I invited all eleven individuals who had published or presented content related to the use or development of rubrics in program evaluation as an evaluation-specific tool by sending each individual a personalized email. I was able to interview seven of the eleven key informants. As the informant group was very specific, I also used a snowball sampling approach by asking participants if they could provide names of other evaluation academics or practitioners who could potentially inform the study. This process resulted in seven more invitations, from which three more interviews were conducted. I interviewed a total of ten individuals, which is a 56 percent response rate, although one individual identified through the snowball sampling technique did not actually use rubrics to draw explicitly evaluative conclusions, so data from that interview was not included in the study. Data from nine interviewees were retained. Research on participation rates of "experts" (e.g., intellectual elites) participating in Web-based surveys show response rates ranging from 32 to 37 percent (Cycyota & Harrison, 2006; Sheehan, 2001). If this holds true for interviewing, the response rate attained is acceptable.

Interviewee Characteristics

Interviewees had an average of 18 years of experience in evaluation, ranging from 8 to 30 years. They had been using rubrics for an average of 12 years, ranging from 7 to 26 years. On average, they each belong to 3 professional evaluation associations or societies, ranging from 1 to 5 organizations. I interviewed both men and women. The highest degree individuals had attained were an undergraduate degree (1 individual); a

post-graduate/master's degree (5 individuals); or a doctoral degree (3 individuals). Interviewees were located in Australia, New Zealand, and the United States.

Interview Protocol

A draft of the semi-structured interview protocol was reviewed by 3 evaluators with 50 years of combined evaluation experience. The protocol was also submitted for HSIRB[1] approval and designated as "approval not needed." The questions were grouped in four main areas: 1) background and training in evaluation; 2) background in rubric use (aimed to elicit specifics on how rubrics are developed and used to draw evaluative conclusions in program evaluation); 3) details on learning to use rubrics; and 4) the importance of learning to use them. Interview data was confidential. To preserve confidentiality, all informants are referred to in this study in the feminine.

Data Analysis

I recorded data by taking notes during the interviews. Note-taking allowed me to engage in a first tier of data transformation by immediately paraphrasing critical elements of each conversation. I reviewed interview notes for accuracy within 24 hours. I sent participants my interview notes to verify that their intent was accurately conveyed. Any follow up clarifications were addressed via email. When I was given consent by participants, I also recorded conversations electronically. This allowed me to refer back to the recordings for clarifications and to extract direct quotations.

I analyzed the notes using an overarching pre-specified approach to ensure alignment with the research question. In other words, because I specifically sought out

[1] Approval not needed for Human Subjects Institutional Review Board (HSIRB) Project Number 14-09-06, Western Michigan University

these individuals for their experience working with rubrics to draw evaluative conclusions, I wanted to be sure to analyze the data to better understand how they do it, how they learned to do it, and how important they believe using rubrics in this way is. Within these broad categories, I used an emergent inductive data analysis technique. I did so by extracting and organizing short sections of the discussions framed by the dimensions of the research questions. My main interest was in the content (actions and events) of the interviews with an eye on, what Elliott (2005) refers to as, the trajectory or advancement of achievement over time, which in the case of rubric use would be progressive mastery. My analysis of the interviews was iterative with new accounts confirming features of previous interviews resulting in a combined account of the nature of the experience for these evaluators (Elliott, 2005). My analysis and my presentation of results purposefully show the similarities between interviews rather than the differences (Labov & Waletzky, 1967/1997). To enhance validity, my presentation of results is supported by interview excerpts to allow the interviewees to "speak for themselves."

Limitations

The validity risks associated with a small sample size are that the perspectives of the sample do not represent the population. In this case, the population is all evaluators who use rubrics as an evaluation-specific methodology, which appears to be a small group of individuals. So, this validity risk is offset because the pool of potential informants is small, thus the number of interviews conducted was also small. The issue is also offset by a reasonable response rate and by the point that the perspectives of participants involved in interviews conducted toward the end of data collection, added little new information, suggesting data saturation.

Results

The results are organized by how interviewees use rubrics, how they learned to use rubrics, and their perspectives on the importance of learning to use rubrics.

How Interviewees Use Rubrics

I wanted a clearer understanding of how evaluators who use rubrics to draw evaluative conclusions do it, in other words, the specifics about their *application* of training or learning. The following describes more specifically: 1) what interviewees use rubrics for; 2) what they believe the characteristics of good rubrics and rubric use are; 3) how they may engage stakeholders in rubric development and use; 4) the contexts in which rubrics work best; and 5) their comments on a counter-factual, i.e., how their evaluations would be different if they didn't use rubrics.

Uses. When asked how they use rubrics, some examples of replies were:

> I use them to make judgments of quality. (I6)
>
> I use them to define criteria and the evaluative reasoning process. (I3)
>
> They are solid and robust valuing frameworks. (I7)
>
> To draw value judgments of quality and success. (I9)

Upon further discussion, it became apparent that these evaluators use rubrics for multiple purposes. They use them to focus and guide evaluation projects. For example, interviewees noted:

> [I use rubrics] every step of the way. You can use a rubric to plan, develop, set up your analysis, and even structure your reporting and dissemination…A good rubric is one that can be totally embedded in the process of evaluation from go to whoa. (I2)

> We use them to identify aspects of the evaluand, the thing we are evaluating. Those aspects of importance and value to the program itself. So it helps focus our evaluation design. (I7)

Even when an evaluation did not require the use of rubrics, some interviewees still developed and used them internally to organize and value study evidence for their own benefit and understanding. As an example, one interviewee said:

> I always go to the evidence about the area I am working in even if they don't want evidence-base[d evaluation] because I need to understand it and a rubric allows me to see what the evidence is saying and what the pathways are for the program that I might be looking at. I will always be making a judgment even if they don't want me to. It is just the nature of being an evaluator. (I2)

Rubrics were also noted to be used at different judgment levels. These are sub-rubrics that apply to the criteria of a core rubric. An example given was an instance when the results of data collection rubrics were applied to a main rubric that was used for a final synthesis, which is an overall judgment of merit. These sub-rubrics were referred to as *nested rubrics* by one interviewee:

> I also use nested rubrics. That happens when there is one overarching rubric for synthesis of program judgment and then data collection rubrics that feed into the overarching rubric. (I1)

Characteristics of good rubrics. While all participants said they use rubrics as a guiding framework, it was noted that rubrics have to be developed to be flexible enough (negotiable) to accommodate unexpected results. Unexpected results could be unintended programmatic consequences, or finding that the rubric was a poor fit for the collected data. Flexibility can be built into a rubric by purposefully adding in a "catchall" criterion. The results grouped into the catchall criterion are later analyzed for important unexpected issues, themes, and consequences that if needed, would be negotiated to fit into the rubric by, for example, expanding it. Another way that flexibility was brought up by interviewees was in reference to setting clear expectations among stakeholders that renegotiation and adaption of the tool are common and often essential, with (participatory evaluation) or without (independent evaluation) stakeholder input, as appropriate. For example, one interviewee said:

> Best practice is to not have criteria locked in, so adjustments can be made – but not in order to compromise. It is an iterative process and when changes are made, it is for the right reasons, for example, because we've learned more about the evaluand, more about what matters, and not to hide an inconvenient truth. (I1)

Interviewees noted that good rubrics have certain qualities, among them clarity and precision. Some examples from the interviewees are:

> The rubric needs to be clear, but not too precise or constrained. (I3)
>
> It depends on the project. Generally I think you need to get the rubric to a high enough level, so there are say five key [criteria] areas and about five standard levels. (I5)
>
> There is a fine balance between rich description and being too wordy. (I6)
>
> A rubric should fit on one page. (I6)

When describing the characteristics of effective rubrics, interviewees said that for rubrics to be practical, they need to be designed to enable use that is iterative, transparent, and participatory.

> A good rubric is transparent and clear with dimensions that come from valid sources, not just decision makers' input. The process needs to include iteration, allowing the rubric to evolve, with refinements recorded over time. (I3)

Good rubrics also need to be realistic.

> [It] needs to be realistic about the levels of performance. The excellent [category] has to be achievable, yet a high standard. It has to reflect all criteria and the movement between levels. (I9)

Participatory development and use. A good rubric development process was described as including the evidence-base (theory, literature, research), but also developed

in a participatory manner to ensure that the unique context of the program is considered. For example, interviewees said:

> You need the right knowledge and context to inform the rubric based on literature and disciplinary and cultural expertise. It should not necessarily be based only on evaluator expertise and stakeholder input. (I3)

> I would probably argue if you are going to build a rubric, you need to bring the evidence-base, but you also then need to understand the evidence-base from what truly happens in practice. And, combine those two. It's a way of validating a rubric. (I2)

More than one participant stated that stakeholders should always be involved in rubric development, but at different levels, as feasible and according to their interest. The point was made that *without* stakeholder input, the evaluation and the judgments are being done *to* them, instead of *with* or *for* them. Stakeholder involvement in rubric development was characterized as an important way to involve stakeholders in an evaluation.

> I think getting the stakeholders' perspective is critical. (I2)

> You need to get the right people and right perspectives together, and people who are willing to invest enough time to develop the rubric properly. (I1)

> The ones that work well have sufficient people at the table at the onset, so key stakeholders have a voice. Without that early participation, we won't have the buy-in at the end. (I7)

Interviewees also spoke of involving stakeholders to enhance credibility.

> Credible means getting the right people in the room who have the right knowledge and expertise. We need to bring research into the rubric and reference two to three sources. It is also important to document who was in the room and their qualifications. This is especially important for high stakes project. (I6)

> We conduct conceptual face validity by engaging with clients and key stakeholders and we adapt the rubric if it isn't right. (I7)

Interviewees made the point that facilitation skills are an important skill when using rubrics in program evaluation.

> You need strong facilitation skills for a good development process. You need to facilitate descriptions of excellent and poor, and discuss the differences—what are the thresholds, what are the deal breakers, what must you have [in the rubric]? You need to facilitate a robust discussion that also includes literature. This secures buy-in and is part of the socialization of the rubric into the organization. It also helps develop an understanding about the whole evaluation process. (I6)

Contexts. When asked to think about specific contexts in which rubrics work best (e.g., types of evaluation approaches, sectors or disciplines, and stakeholders) interviewees said rubrics could be used in any context.

> They just work. Giving explicit evaluative reasoning in rubrics works well. (I1)
>
> Rubrics can be used in any context. If it is desk research, just develop the rubric on the literature. (I5)
>
> Rubrics can work anywhere, but based on communication of purpose, utility, and value-added to the organization—articulating the benefits. Rubrics take energy, effort, and time. We need to show the value and recognize the driving forces in different settings to do this. In some settings, I use culturally understood metaphors to explain the value; in other settings I go into how it makes business sense. (I6)

Some of the warnings about rubric use were that they shouldn't be used as just a "tick-the-box exercise" (I3). Nor can they be popped in from one project to another (I8). Rubric fatigue was also brought up as an issue, which can develop from rubrics that are "overdone, or done poorly" (I5).

The counter-factual. The interviewees said if they had never used rubrics, their evaluations or practice would be "a mess" (I2), "stumbling" (I1), "wordy" (I4), "descriptive" (I3), "research" (I5), "bloody terrible" (I6), "clumsy and unethical" (I7), "not as sharp" (I8), and "bland—it would lack valuing" (I9).

> I would still be stumbling along thinking I was doing good evaluation and having the same problems I was having before—not being able to arrive at or articulate how I reached a judgment, if I was even able to reach a judgment. People who want to question or challenge a report—it leaves a

> great risk of appearing subjective if evaluative judgment is not explicit. (I1)
>
> I would be making judgments on a whim. Rubrics frame how we collect data. They engage stakeholders. I am more inclusive and collaborative due to rubric use. (I6)
>
> I would not be sufficiently explicit about values. (I5)
>
> Before my findings were very descriptive, a kind of connoisseurship summing up…to make judgments. I was answering questions which were not always evaluative. (I3)

Overall, interviewees expressed that they use rubrics to draw evaluative conclusions, but that they also use rubrics in a broader manner, to frame their evaluations. Good rubrics are developed to discern between specific criteria and levels (around 5 or 6 of each); have the right mix of description without being too wordy; and have built-in flexibility. Interviewees communicated that stakeholder input is critical, but that their input needs to be supported by empirical evidence and literature. Interviewees also related that they feel they can use rubrics in any context, and that their practice is substantially better when using rubrics, as opposed to when not using them.

Learning to Develop and Use Rubrics

Information about how interviewees learned to develop and use rubrics is organized below with regard to their 1) initial encounters; 2) catalysts for use; and 3) journeys to competence.

Initial encounters. None of the participants interviewed had received rubric training in their degree programs. Instead, they accessed information in two ways: 1) as trained educators, they learned how to use rubrics; and/or 2) having contact with Jane Davidson and her body of work.

The two interviewees who hold degrees in education related in the interviews that the use of rubrics was a natural part of an educator's practice and transitioning rubrics over to program evaluation was also quite natural. For example, when asked how long she had been using rubrics in her practice, one interviewee with a background in education replied:

> Probably forever. I'm a trained teacher and very much involved in understanding assessment as a consequence of my training as an educator and also teaching teachers myself. I think rubrics have been a foundation of assessment for many years. So as a teacher, I've been able to use the ideas about assessment in my evaluation work. For example, I do a lot of standard setting in evaluation. To do appropriate standard setting, you need to develop rubrics. So, probably since 1989 I've been using standard setting, which equates to developing rubrics. (I2)

All other interviewees said they initially learned how to use rubrics through contact with Jane Davidson herself and/or her writing. All but one of the interviewees spoke of being introduced to rubrics by reading Davidson's 2005 book, *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation;* interacting with her at conferences (as fellow presenters or as attendees); attending her workshops; and/or being mentored by her personally, as the following quotes relay:

> I learned how to use rubrics by contacting the person who wrote the book and bought some of her time. (I1)

> First I read Jane's book. This was at the point I was thinking of a career shift. It was a revelation to learn that we could tell clients if a result was good or bad. (I5)

Catalysts. For some interviewees, their initial contact with rubrics brought about an epiphany when they recognized there were deficits in their practice that rubrics helped to fill. One interviewee's epiphany was related to a client's dissatisfaction, which spurred change in her practice. She explained:

> It took multiple exposures to rubrics to convince me to see the value of them. The first was working with a colleague I really respected on a three-year process and outcome evaluation. She said I needed to explicitly add in criteria of merit. I had never heard the phrase "criteria of merit" before. She gave me Jane's book. I was interested and tried it in the three-year project, in stages: Year 1, going into field to see what matters to the people; then I developed rubrics in the second year; and then I made judgments based on the rubrics in the third year. It went ok. Another exposure was when I was a keynote speaker in a conference. Jane was also a keynote. I heard her speak on rubrics and had time to interact with her after. Then I was hit between the eyes. In another evaluation, I presented results to stakeholders in an effort to get feedback on the report. I had not used rubrics on this project. A psychiatrist was in the audience and

> basically said the conclusions were "crap"—he was saying there was no clear sense of what I was basing my evaluative judgments on. So, I thought: OK—time to really master this rubrics thing! (I1)

Another interviewee intertwined her understanding of the logic of evaluation (developing criteria, standard setting, and importance weighting) and rubric use. This person noted exposure to these two areas as the catalyst that moved her evaluation practice from an applied research approach to a valuing approach—an approach that combines evidence with values to draw explicitly evaluative conclusions. For others, it was a great relief that such a tool exists. She noted:

> Finding rubrics was just like magic, because for someone coming out of [a different professional field] it was like Yes! Now it's easier to nail the focus of the work and the data. Now it's easy to figure out what questions we need to ask. Now it's easy to incorporate other data sources. We can incorporate the literature. Because we're going to map the whole lot against the criteria. It was a Eureka moment for me. And that's why I persevere with them because even if it takes the client a bit longer at the start sometimes, I just find it makes the rest of the process so much smoother. And I just stack all my data, when I'm doing my analysis, against the criteria and then it becomes my analysis tool. (I5)

Journey to competence. As I listened to interviewees relay how their thinking about rubrics evolved over time, it became apparent they were on a rubric-learning journey.

> I started with easy rubrics, lazy rubrics. They were generic with very broad criteria. I am not a big fan of doing this anymore because it postpones the hard work—at the end you still have to explicate how you got to the judgment(s) (I1).

> When just starting, they take a long time to develop…. My first rubrics were maybe too comprehensive. Not the best. I was figuring out how to bring data to bear with the rubric and getting a grip on this in a structured way. And learning to fine tune judgments. (I6)

For many, the journey seemed to begin with a novice approach of trying to include everything possible in a single rubric. One interviewee said it was probably a way of overcompensating for lack of experience.

> In my first attempt, I had 6 rubrics spanning 25 pages. It was very long and I threw in the kitchen sink. Not great, but it was a start. (I1)

> Learning to do rubrics was quite hard work. Our first one was huge! We had 9 evaluative criteria and 53 sub domains within that. So you do learn along the way. (I5)

Strategies that the interviewees described for becoming competent in rubric development included practicing using them (just doing it), engaging in a community of practice, being mentored, and seeking out professional development.

> [I recommend] learning by doing. There are more people using them now, so working with others helps too. Mentoring in with Jane was good, and I also recommend reading Jane's book. (I1)

> Books and conference presentations can be a starter—but to really learn to do them, you need mentoring. (I5)

For many on the journey, it seemed fine for them to be on a seemingly unending quest as they revel in the emergence of new strategies and ways of thinking about rubrics.

> I was presenting a conference session with a colleague and when I asked her about her presentation, she said, "I got the slides from you last year." When I looked them over, I realized that she was presenting an old way that I used to think. That I had moved on from that understanding. (I5)

They often call on each other for support.

> We have a unique community of practice where we are a group of independent consultants who call on each other all the time, especially when there are sticky problems. (I5)

They work with others who are just beginning the journey (seven of the nine interviewees had engaged in some type of training of others—conference sessions, workshops, and mentoring).

> [I train others] all the time. There are phases of a career, and in this phase, I work as an expert advisor and I can try to make it count [for others]. (I1)

They also share their experiences with those who would debate the value of the journey.

> I engaged in a really interesting conversation with [a visiting scholar] who professed to not being a fan of rubrics but after talking together, this person admitted to seeing the usefulness of them. (I5)

In summary, I found that learning to use rubrics is a journey. Most interviewees initially learned to develop and use rubrics through contact with Jane Davidson and her body of work. Many latched on to them because they saw that rubrics filled gaps in competency such as pushing practice from applied research (shrouded as evaluation) to e*valu*ation (practice that actually values).

Importance of Using and Learning to Use Rubrics

As a proxy indicator for importance, I asked interviewees how often they use rubrics when conducting evaluations in their practice. They noted using rubrics for evaluations a median of 99 percent of the time, ranging from 100 percent to 50 percent.

> I use rubrics now in 90 percent of my practice, but when I don't use them, I still always present explicit evidence about judgments. (I1)

> Most projects involve rubrics. We develop the key evaluation questions and then the rubrics. (I4)

> They are the core of the evaluation-specific methodology. I can't imagine doing an evaluation without them because the valuing piece would be

> missing – values would not be explicit and nailed down at the start, and it would be harder to get the project going and keep it on track. (I5)

I asked interviewees to comment on how important they think rubric training is for evaluators. The overwhelming sentiment was that it is critical.

> Really important. It is an integral part of doing transparent evaluation. (I4)

> Because rubrics provide the basis for evaluation (as a tool for practice), they should be a part of professional development and degree programs in evaluation. (I6)

> I think more training on criteria and rubrics that fit with the criteria and standards is needed. How we deal with criteria per se needs to be more systematic, so we need training to support that. (I3)

> It is a vital tool in the toolkit. Evaluators need to know how to use them. (I7)

> Hard to think of anything more important. It shows explicit evaluative reasoning. If an evaluator is not using rubrics, how are they doing explicit evaluation well? How are they showing how they came to their conclusions? It needs to be a part of any evaluation course. How can you be an evaluator if you aren't getting trained in what Scriven called the "fundamental problem in evaluation," which is how you get from empirical evidence to explicitly evaluative conclusions. The way you do that is through criterial inference, which is a specific type of probative

> inference. You can do this type of inference with or without rubrics, but it is a hell of a lot easier with rubrics. (I1)
>
> I think it would be great if there was training in the use of rubrics, but it's quite hard to do because there is the theory, and people generally get the theory quite quickly, but evaluators also need the skills that experience and maturity bring, to navigate use and buy in with clients. It's applying rubrics in practice that is where it gets quite challenging – and it's the intersection of evaluative thinking, client management and project management – that praxis piece that is so hard. (I5)
>
> It is vital. [Before I learned to use rubrics,] I used to be debating with clients at the end of an evaluation about the measures used. But looking back, they were really talking about the judgments and it just came out in debates about measures. Now that I use rubrics, those debates don't occur anymore. (I6)

Conclusions

Interviewees were chosen for their experience using rubrics to draw transparent and explicitly evaluative conclusions. Beyond this purpose, evaluators also spoke of using rubrics more broadly as frameworks or guides to sharpen an evaluation's focus. They use rubrics to develop shared understandings among key stakeholders and to negotiate with stakeholders about what *good* looks like, in theory and in context. They also use rubrics to guide discussions with key stakeholders after judgments have been made to make sense of those judgments.

Interviewees overwhelming noted that using rubrics is an important aspect of their evaluation practice and that training to use rubrics is critical for evaluators. By definition, to conduct evaluation, evaluative determinations must be made, but there is little guidance on how to synthesize evidence and values to make the determinations. Rubrics are tools that fit that purpose, but again, there is little information in books, peer-reviewed literature, or professional development materials about how to develop and use rubrics for reaching explicitly evaluative conclusions. None of the interviewees had formal training in how to develop and use rubrics for program evaluation. The main way that interviewees learned to use rubrics was through some type of engagement with Jane Davidson and her body of work. These interviewees are cutting-edge, early adopters, with many honing their skill in rubric use by being part of a rubric community of practice. In fact, while currently located in four countries, every interviewee lived and worked in New Zealand over the past decade—a period of time that Davidson was also in New Zealand and a central part of this community of practice. Having such a strong champion to promote rubric use in program evaluation allowed the practice to gain traction, but logically, in the long run this training strategy is not efficient and is not sustainable.

By definition, drawing evaluative conclusions is a core function of evaluation (AEA, 2014). However, doing so is a complex task, which may be why Davidson (2013) asserted that providing evidence without drawing it together for evaluative conclusions to determine something's merit, worth, value, or significance, may be left undone by evaluators. Arens' (2005) findings that award-winning evaluations rely heavily on implicit reasoning, also supports this notion of the difficulty of making evaluative

synthesis explicit. My research adds to and supports these findings. For example, when interviewees were asked if they had *never learned* to use rubrics, how they would achieve this function, they all replied that they would reach evaluative conclusions by providing descriptive narrative. They also characterized their evaluations without rubrics as being "a mess" (I2), "stumbling" (I1), "wordy" (I4), "descriptive" (I3), "research" (I5), "bloody terrible" (I6), "clumsy and unethical" (I7), "not as sharp" (I8), and "bland—it would lack valuing" (I9). Rubrics are used to combine evidence and values, whilst also making the evaluative reasoning that Arens (2005) found to be implicit, explicit. My research suggests that rubric use assists in the difficult task of drawing explicitly evaluative conclusions, and that alternative strategies for doing so are limited, at least within this interviewee group.

Revealing the need and gap in evaluator training around this tool, a tool so critically important to the few who have been able to learn how to develop and use rubrics, and make them integral to their evaluation practice, is the first step in driving future focus in formal and informal evaluator training. These early adopters are honing their skills through trial and error and when possible by tapping into a community of practice. To expand understanding, application, and acceptance of use, systematic training on rubric development and use in program evaluation needs to be undertaken.

Future research should focus on investigating proven practices for developing and using rubrics and how rubric use affects evaluation quality or use. Such research can form a foundation for development of further guidance, training, and resources.

References

References marked with an asterisk indicate evaluation books used to identify interviewees.

*Alkin, M. C. (2011). *Evaluation essentials: From A to Z*. New York, NY: Guilford.

American Evaluation Association. (2014, January 10). *What is Evaluation?* Retrieved from http://www.eval.org/p/bl/et/blogid=2&blogaid=4

Arens, S. A. (2005). *A study of evaluative reasoning in evaluative studies judged "outstanding"* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (304986745).

*Bamberger, M., Rugh, J., & Mabry, L. (2011). *Real world evaluation: Working under budget, time, data, and political constraints* (2nd ed.). Thousand Oaks, CA: Sage.

*Chen, H. T. (2014). *Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, & mixed methods approaches* (4th ed.). Los Angeles, CA: Sage.

Cycyota, C. S., & Harrison, D. A. (2006). What (not) to expect when surveying executives: A meta-analysis of top manager response rates and techniques over time. *Organizational Research Methods, 9*(2), 133-160.

*Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davidson, E. J. (2013). *Actionable evaluation basics: Getting succinct answers to the most important questions.* Auckland, New Zealand: Real Evaluation.

Dirlam, D. (2015). Changing wisdoms: History of developmental rubrics. Ed. David K. Dirlam. Retrieved from http://changingwisdoms.org/?s=Developmental %20Rubrics&sid=2&ss=History&ssid=1

Elliott, J. (2005). *Using narrative in social research: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

*Fetterman, D. M. (2001). *Foundations of empowerment evaluation.* Thousand Oaks, CA: Sage.

*Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2010). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Boston, MA: Pearson.

Labov, W., & Waletzky, J. *(*1967/1997*).* Narrative analysis: Oral versions of personal experience. Reprinted in *The Journal of Narrative and Life History, 7*, 3-39.

Martens, K. S. R. (2015). *Rubrics in program evaluation*. Unpublished manuscript.

*McDavid, J. C., Huse, I., & Hawthorne, L. R. L. (2013). *Program evaluation and performance measurement: An introduction to practice* (2nd ed.). Thousand Oaks, CA: Sage.

Mertens, D. M., & Wilson, A. T. (2012). Program evaluation theory and practice: A comprehensive guide.* New York, NY: Guilford.

*Owen, J. M. (2006). *Program evaluation: Forms and approaches* (3rd ed.). Crow's Nest, New South Wales, Australia: Allen & Unwin.

*Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Los Angeles: Sage.

*Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford.

*Patton, M. Q. (2011). *Essentials of utilization-focused evaluation.* Thousand Oaks, CA: Sage.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership, 55*(2), 72-75.

*Posovac, E. J., & Carey, R. G. (2011). *Program evaluation: Methods and case studies* (8th ed.). New York, NY: Routledge.

Reddy, Y. M., & Andrade, H. (2014). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.

Research. (2015). *Merriam-Webster Dictionary*. Retrieved from http://www.merriam-webster.com/dictionary/research

*Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

*Royse, D., & Thyer, B. A. (2016). *Program evaluation: An introduction to an evidence-based approach.* (6th ed.). Boston, MA: Cengage Learning.

*Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Sheehan, K. B. (2001). E-mail survey response rates: A review. *Journal of Computer-Mediated Communication, 6*(2), n.p.

*Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). San Francisco, CA: Jossey-Bass.

USAID Center for Development Information and Evaluation. (1996). *Performance Monitoring and Evaluation TIPS*. Number 2. Washington, DC. Retrieved from http://pdf.usaid.gov/pdf_docs/pnabs541.pdf

*Weiss, C. H. (1997). *Evaluation: Methods for studying programs and policies* (3rd ed.). San Francisco, CA: Jossey-Bass.

*Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (2010). *Handbook of practical program evaluation* (3rd ed.). San Francisco, CA: Jossey-Bass.

*Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

IV. HOW THE FORM AND FUNCTION OF RUBRICS MAKE EVALUATIVE REASONING EXPLICIT

Introduction

"Understanding the reasoning process to establish evaluative conclusions drawn in practice has to be the field's greatest unmet challenge" (Fournier, 1995, p. 1). This study presents how the form and function of rubrics can make evaluative reasoning explicit. The American Evaluation Association (AEA) defines evaluation as "a systematic process to determine merit, worth, value or significance" (2014). If written reports include conclusions and recommendations, whether or not evaluative claims are made, the fact that conclusions are specified and recommendations are given, denotes that at some level, evaluative conclusions have been reached (e.g., suggestions for improvement are given which means that something is *not good enough* and *in need of fixing*). To make such determinations, one has to have some set of standards to guide reasoning (Davidson, 2005). Arens (2005) found in her study of Outstanding Evaluations, as awarded by the American Evaluation Association, that the conclusions and recommendations of the five studies she examined relied heavily on implicit criteria and lacked specifics about the synthesis process upon which judgments contained in the conclusions and recommendations were based.

Professional fields have common standards and competencies to guide practice. The Program Evaluation Standards (PES) provide guidance for evaluation practitioners and evaluation consumers (Yarbrough, Shulha, Hopson, & Caruthers, 2011). The PES are organized into five categories: Utility; Feasibility; Propriety; Accuracy; and Evaluation Accountability. Among guidance provided by the PES are standards for transparency and

disclosure (Propriety Standard 5) and explicit evaluative reasoning (Accuracy Standard 7). Similarly, evaluator competency frameworks "are designed for everybody who wants to enhance their effectiveness as an evaluator or commissioner of evaluations" (AES, 2013). Evaluator competency[1] frameworks developed by the Australasian Evaluation Society (AES, 2013) and the Canadian Evaluation Society (CES, 2010) explicate competent evaluation practice. According to the CES, applying the Program Evaluation Standards is a key evaluator competency (CES 1.1); thus, competent evaluators would apply PES standards P5 and A7 outlined above. Providing answers to evaluation questions (CES, 2.15.1), and providing clients with value judgments (AES, 2.4.1) are also key evaluator competencies. Thus, drawing transparent and explicit evaluative conclusions to answer evaluation questions is a core competency for evaluators. This is an aspect found lacking in award-winning evaluations. If the interpretation step of award-winning evaluations is not transparent and explicit, others are unable to learn from these exemplary practitioners, and the lack of transparency and explicitness also hinders replicability, diminishing validity and feeding into the pervasive notion that evaluation is subjective. Taken one step further, if *award-winning evaluations* do not include transparent explicit reasoning, the problem is likely to be widespread throughout evaluation practice.

One solution to this problem is for evaluators to use rubrics. Rubrics are tools that transparently communicate the reasoning behind systematic syntheses to draw defensible evaluative conclusions (Davidson, 2014; King, McKegg, Oakden, & Wehipeihana,

[1] These two sets of competencies were chosen because they are the most relevant to the author's evaluation practice (US and Australia) and they have been ratified by their relative evaluation societies. Unfortunately, the American Evaluation Association has yet to formally adopt a national set of evaluator competencies, so the Canadian set is included because it is the set most commonly deferred to by US evaluators.

2013). Rubrics have been used in education since the 1970s to evaluate student performance by scoring and/or grading artifacts of academic work (Dirlam & Byrne, 1978). However, a thorough review of peer-reviewed program evaluation literature produced few examples of the use or described utility of rubrics to reach explicitly evaluative conclusions (Martens, 2015a). Interviews with practitioners who use rubrics in program evaluation found rubrics were useful in multiple ways, importantly, including using them to reach explicitly evaluative conclusions, but such practitioners rarely publish their experiences in the peer-reviewed literature (Martens, 2015b). The aim of this study is to explain how the form (characteristics and configuration) and function (the natural purpose) of rubrics exemplify the core logic and nature of evaluation and in so doing can make evaluative reasoning explicit (Scriven, 1981; 1991). This is accomplished by 1) positioning rubrics in evaluation logic and theory, and 2) describing the components of a rubric as they relate to the nature and logic of evaluation, as conceptualized by Scriven (1981) and Fournier (1995).

Positioning Rubrics in Evaluation Logic and Theory

In his book, *Evaluation Thesaurus*, Scriven wrote, "The key function of evaluative inferences is moving validly to evaluative conclusions from factual (and of course definitional) premises; so the key task of *the logic of evaluation* is to show how this can be justified" (emphasis added, 1991, p. 216). Fournier (1995) summarized Scriven's seminal work, *Logic of Evaluation* (1981), into four steps: 1) establish criteria; 2) construct standards; 3) measure performance and compare with the standards; and 4) synthesize evidence into a judgment (of merit, worth, or significance).

The nature of evaluation is fundamentally how evaluative conclusions can be derived; it is how Fournier's fourth step is possible. Metaphorically, Scriven characterizes *the nature of evaluation* as having two arms. One arm is engaged in data gathering (data that is relevant to the criteria of the evaluand) while the other arm is engaged in identifying relevant values and standards. Relevant values in evaluation are the aspects of the evaluand that merit is based upon and are drawn from sources such as a needs assessment, literature, professional standards, or a logical analysis of the function of the evaluand—they are not individually held values (Davidson, 2005; Scriven, 1991). Scriven describes evaluative synthesis as the "head" that synthesizes the results of these two types of activities. Evaluative synthesis is combining factual premises relevant to the criteria with "evaluative premises or standards" to make an inference that is systematic and defensible (1991, p. 5). Scriven writes that one of the ways to make such an inference is through the *logic of descriptors* "which are abstractions from a complex of indicators" (1991, p. 218). Three key words run through the above discussion: 1) criteria; 2) standards; and 3) descriptors. These three terms are important because they are the fundamental components of a rubric. Criteria "are the aspects of an evaluand that define whether it is good or bad and whether it is valuable or not valuable" (Davidson, 2005, p. 239); standards are the spectrum of goodness by which criteria are judged; and descriptors describe aspects of performance for each criterion at each standard level. Figure 14 shows an example of how these three components can be combined in a rubric, although the placement of the criteria (a) and the standards (b) could just as easily be swapped for the same effect. The configuration can be decided according to the best fit for presentation.

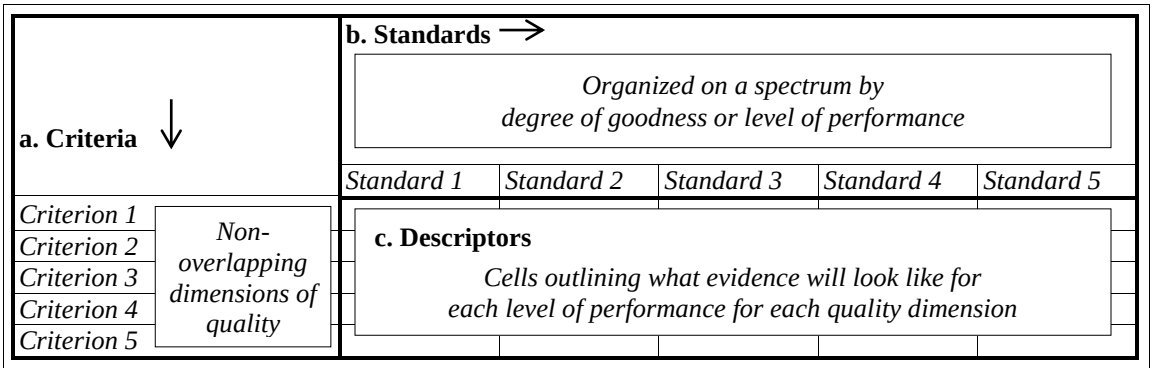


Figure 14. Three components of an evaluation-focused rubric: (a) Criteria, (b) Standards, and (c) Descriptors. From "Rubrics in Program Evaluation," by K. S. R. Martens, 2015a, unpublished manuscript, p. 14. Reprinted with permission.

Figure 15 is an example rubric that illustrates criteria (left column), standards (4, 3, 2, 1), and descriptors (interior cells).

Evaluation Rubric for IPOD APPS

Criteria ↓	4	3	2	1
Curriculum Connection	Skills reinforced are strongly connected to the targeted concept	Skills reinforced are related to the targeted concept	Skills reinforced are prerequisite or foundational skills for the targeted concept	Skills reinforced in the app are not clearly connected to the targeted concept
Authenticity	Targeted skills are practiced in an authentic format/ problem-based learning environment	Some aspects of the app are presented in an authentic learning environment	Skills are practiced in a contrived game/simulation format	Skills are practiced in a rote or isolated fashion (e.g., flashcards)
Feedback	Feedback is specific and results in improved student performance; Data is available to the student and teacher	Feedback specific and results in improved student performance (may include tutorial aids)	Feedback is limited to the correctness of student responses and may allow students to try again	Feedback is limited to the correctness of student responses
Differentiation	App offers complete flexibility to alter settings to meet student needs	App offers more than one degree of flexibility to adjust settings	App offers limited flexibility (e.g., few levels, such as: easy, medium, hard)	App offers no flexibility (settings cannot be altered)
User Friendliness	Students can launch and navigate within the app independently	Students need to have the teacher review how to use the app	Students need teacher to review how to use the app on more than one occasion	Students need constant teacher supervision in order to use the app
Student Motivation	Students are highly motivated to use app and select it as their choice among related choices of apps	Students will use the app as directed by the teacher	Students view app as "more school work" and may be off-task when directed by the teacher to use the app	Students avoid use of the app or complain when use of the app is assigned by the teacher

Figure 15. Example rubric. Adapted from "Evaluation Rubric for IPOD APPS," H. Walker, 2010, *John Hopkins University*. Reprinted with permission.

Since it is an evaluator's imperative to draw evaluative conclusions, and since rubrics embody the nature and logic of evaluation, discussing how rubric use fits with varying evaluation approaches may broaden their acceptance. Marvin Alkin's (2012) book *Evaluation Roots* presents three main branches of "the evaluation theory tree": 1) the valuing branch; 2) the use branch; and 3) the methods branch.

The *valuing branch* includes evaluation approaches that prioritize making judgments of merit, worth, or significance. A rubric is an obvious fit into the toolbox of evaluators who adhere to these approaches as it aids in the task of combining evidence with values to draw evaluative conclusions.

The *use branch* includes approaches that focus on the intended use of evaluation results by evaluation users. Evaluators who subscribe to approaches in the use branch should see the benefit of employment of rubrics as rubrics explicitly reveal how judgments of programs' merit, worth, or significance are made. This practice promotes user understanding and buy-in, especially if the rubric is developed in a participatory manner, aspects foundational to the use branch (Patton, 2008).

The *methods branch* of the evaluation theory tree includes theorists who focus on optimizing study rigor (Alkin, 2012; Cardin & Alkin, 2012). Evaluators whose approaches fit on the methods branch would be concerned about the validity and reliability of using rubrics. Rubrics are not unlike employing other instruments (e.g., surveys or interview protocols); they must be carefully developed and refined to meet the standards of rigor required within the context of an evaluation.

Essentially, the core components of a rubric incorporate elements of the nature and logic of evaluation (criteria, standards, and descriptors). While not all evaluations

require rubrics, all evaluations do require evaluative reasoning to accomplish the task of drawing transparent and explicit evaluative conclusions to answer evaluation questions. Drawing transparent and explicit evaluative conclusions is a core competency for evaluators and a rubric is a very useful tool for this purpose (Davidson, 2014) that can be used with any evaluation approach. The following section describes and analyzes the components of a rubric based on Scriven's logic and nature of evaluation (1981; 1991).

Rubric Components

The components of a rubric are examined below and are organized by: 1) criteria; 2) standards; and 3) descriptors.

Criteria

In evaluation, criteria are what matter in the program, what counts, the big important things (Davidson, 2005). Criteria are the aspects we look at covering: need; design; implementation; outcomes; cost-effectiveness; to determine how good something is (Davidson, 2005). A rubric connects the focus developed in the evaluation design stage to the evaluative determinations that will be concluded at the end of the study. Based on the language of the *Logic of Evaluation* (Scriven, 1981) the evaluation design elements that play into criteria development are as follows: 1) the evaluation predicates (merit, worth, significance); 2) the perspective of the evaluative conclusions as holistic or analytic; and 3) the lexicon of the evaluative determination (generic or specific).

Merit, worth, significance. The evaluation conclusions (Fournier's Step 4) are dictated by the basic evaluation predicate(s) that are at play in the study. The predicate(s)

are established in the design phase of an evaluation and as such are reflected in the evaluation questions. The evaluation predicates are as follows: merit, worth, and significance (Scriven, 1981). If the evaluation is designed to find out about the quality of a program, then the evaluation questions will focus on determining program *merit,* and the criteria in the rubric must be indicative of quality. A list of criteria that outline what makes for a "good" program of this type will be developed and shown in the rubric. If the design and questions are focused on determining how valuable or worthwhile the program is, then the evaluation predicate is *worth* and cost-effectiveness or value for money criteria will be developed (e.g., criteria about things like non-monetary resources invested, monetary investment, and lives affected). If the evaluation design and questions are seeking to determine the importance of a program, then the criteria will focus on the *significance* of the program (e.g., potential influence). Focusing on one specific evaluation predicate does not preclude intentionally using either or both of the other two in the same evaluation or rubric. Table 4 presents examples of criteria of merit, worth, and significance.

Table 4.

Examples of Differences in Criteria by Evaluation Predicate

Criteria of Merit[2]	Criteria of Worth	Criteria of Significance[3]
How good is the program, or are key aspects of the program?	*How cost-effective is the program, or are key aspects of the program?*	*How important is the program, or are key aspects of the program?*
Relevance	Monetary investment	Potential influence
Effectiveness	Person hours	Importance
Efficiency	Materials and resources	Visibility
Sustainability	Infrastructure	Generalizability
Impact	Impact	Impact

Holistic versus analytic. Once settled upon the determination (e.g., merit), the next element of the nature and logic of evaluation that affects the criteria is whether the overall evaluation perspective is holistic or analytic. Scriven (1991) defines *holistic* as synonymous with *global*, which is evaluating the overall character or performance of an evaluand; in contrast to an analytic evaluation which breaks down the evaluand into components or dimensions. The Meriam-Webster dictionary (2015) defines *holistic* as "relating to or concerned with wholes or with complete systems rather than the analysis of, treatment of, or dissection into parts." This translates to rubrics used in evaluation because *holistic* rubrics support broader judgments of performance or quality in a global manner (Moskal, 2000). The key distinction is that holistic rubrics are not tools to synthesize analyzed data; they instead, are tools used to determine the merit of one intact thing without breaking it down into parts and putting it back together. Alternatively,

[2] Listed criteria were developed by the Organization for Economic Cooperation and Development (OECD) Development Assistance Committee (DAC) for evaluating development assistance programs (2015).
[3] Based on Stufflebeam & Coryn (2014, p. 13).

analytic rubrics are for separately evaluating distinct aspects of an evaluand, which can be synthesized for an overall determination if the evaluation design calls for it.

The nature and logic of evaluation outlines that *analytic* criteria are for evaluating programs according to their *dimensions* or *components* and if need be, can be synthesized to make a judgment of overall merit. A *dimension* is an aspect or feature of an evaluand, such as relevance or cost-effectiveness. *Components* are constituent parts of a larger whole (Scriven, 1991). For example, when evaluating a U.S. federally supported Whole School, Whole Community, Whole Child (WSCC) initiative, a *component* of the initiative is the *school health education* program. The other eight *component* programs of the initiative are *physical education*; *nutrition services*; *health services*; *physical environment*; *emotional environment*; *staff wellness services*; *counseling, psychological and social services*; *community involvement*; and *family engagement* (CDC, 2015). Alternatively, the *dimensions* of a WSCC initiative could include areas such as *shared vision, leadership, partnering, responsiveness,* and *impact.* Whether components or dimensions are used, an important consideration when using analytic criteria in a rubric is that the criteria do not overlap (Moskal, 2000). This is a validity and reliability concern because if criteria overlap, raters will have to make decisions about which category to select, which affects reliability (Stemler, 2004) or leads to "double counting" which is penalizing or rewarding in more than one criterion area for the same concern—which affects validity (Moskal, 2000; Scriven, 2000, revised 2005, 2007 p. 3).

Another important consideration when developing analytic criteria is that the criteria are all on the same level or are commensurate (Scriven, 2000, revised 2005, 2007). Being equal in level does not mean being of equal importance within the level;

often criteria are not. Differing levels of importance can be dealt with *within* the rubric through importance weighting if using numeric scores as the standard (Popham, 1997), or *outside* of the rubric through an importance weighting process such as engaging in a qualitative weight and sum procedure (Davidson, 2005; Scriven, 1991). Importance weighting within a rubric occurs when the evaluand is "given performance scores on a standard scale (e.g., 1-10), for each dimension, and the sum of the products of the weights (of each criterion) by the performance (on that dimension) for each [evaluand] is used as the measure of merit" (Scriven, 2000, revised 2005, 2007, p.3). For more information on qualitative weight and sum procedures, see Davidson (2005) and Scriven (1991).

The list of analytic criteria should be clear and complete, yet concise (Scriven, 2000, revised 2005, 2007). If at all possible, there should not be more than five or six criteria in a rubric or it gets too unwieldy and complicated, causing more issues with validity and reliability (Martens, 2015b; Popham, 1997). It is good practice to build a catchall criterion into an analytic rubric to handle emergent unexpected evidence (Martens, 2015b). This evidence can be re-examined during the final stags to determine if the catchall evidence clusters into a theme that makes up a missed criterion that should be added, or if the evidence should be handled by making decisions for placement within the existing criteria.

Analytic criteria, originally termed Standardized Developmental Ratings in the 1970s (Dirlam & Byrne, 1978), are well-positioned to provide information for improvement because, by definition, they are already broken down into parts and the judgments made about the parts can inform on strengths and weaknesses by component or dimension (especially if the descriptors are qualitative), which can then be synthesized

into a higher level determinations, as needed. The global view, however, (using one holistic criterion) is well-suited to providing overarching judgments of merit (Moskal, 2000). A holistic criterion is often used when the criteria of merit overlap or are difficult to tease apart because they are so interdependent (Moskal, 2000). An example from education could be marking student essays (Popham, 1997). In program evaluation, an example could be evaluations of program partnerships. In both examples, if breaking down the complex system of interdependency would negate the potential synergistic effects of the system as a whole, a holistic criterion should be considered. Some rubrics present both holistic and analytic perspectives in the same matrix (Martens, 2015a).

Generic versus specific. Choosing to develop *generic* criteria or *specific* criteria is another rubric development decision. Generic rubrics are, as their name implies, developed to be purposefully broad to allow use in a variety of situations, whereas specific rubrics are developed for use in a limited and distinct context (Davidson, 2005; Moskal, 2000). The criterion or criteria in a specific rubric are tailored to the particular dimension or components of the program being evaluated. Although, Popham (1997) cautions against developing criteria in educational settings that are so specific that they judge the task, and not the higher level skill. Bridging into evaluation, this issue would translate to judging the minutia of indicators instead of getting to the criterial level (Davidson, 2013).

Figure 16 depicts a flowchart of *criteria* characteristics to be considered for rubric development. The figure is meant to lead the reader through decisions from top to bottom starting with whether merit, worth, or significance will be determined; then, if the rubric will contain holistic or analytic criteria, and if analytic, determine if dimensions or

components of criteria will be outlined; and finally determine if the criteria are generic or specific. The figure maintains a basic rubric format as a reminder of the other two rubric components (standards and descriptors) discussed in the following sections.

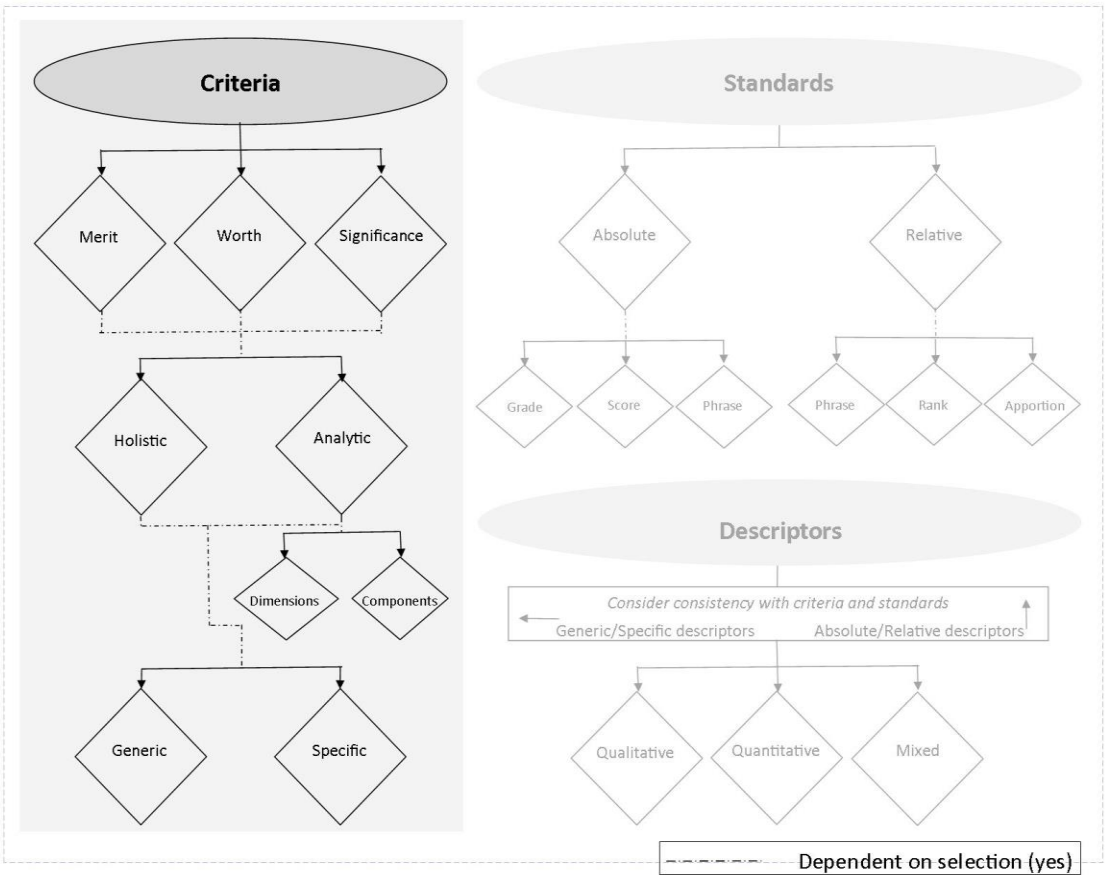


Figure 16. Flowchart of decisions when developing criteria for a rubric.

Figure 17 presents examples of criteria of merit that could be used in evaluations of international development programs. The examples are of analytic criteria split into generic and specific criteria.

Generic (e.g., DAC Criteria, 2015)	**Specific** (e.g., HPI example)
Relevance	*Meeting basic needs*
Effectiveness	*Livestock care and management*
Efficiency	*Environment care and management*
Impact	*Education for a just and sustainable world*
	Empowerment of family and community
Sustainability	*Systems and policy improvements*

Figure 17. Examples of analytic criteria. The criteria should be read vertically by column, the rows are not meant to denote horizontal equivalence. HPI is the acronym for Heifer Project International. The criteria were outlined in an article by Clements, P. (2012).

Standards

Standards make up the spectrum of goodness on which performance is rated. Standards can be expressed in a multitude of ways, including qualitatively, numerically, or as letter grades (Scriven, 1991). The way the standard spectrum is expressed is not limited to one choice; for example, standards can be expressed as both scores (e.g., 0-3) and phrased in explicitly evaluative terms or phrases (e.g., excellent, good, adequate, poor). Options for standards are constrained by the decision to present standards in absolute or relative terms, which is dependent on the type of the evaluative conclusions being drawn (Davidson, 2005).

Absolute standards. *Absolute* standards lay out a spectrum that denotes unqualified goodness or stand-alone merit. Examples of absolute standard ratings are scores, grades, and descriptive labels (Davidson, 2005; Scriven, 1981). Scoring is done numerically and denotes equidistance, so thinking through the appropriateness of using ordinal scales such as Stage 1, Stage 2, Stage 3 is important as it can cause confusion due to the use of numbers which imply equidistance. If equidistant placement of the standards is not intended, it is better to choose letters grades or textual ratings (Scriven, 1991).

Ratings often include phrasing of met need (e.g., met, partially met, unmet) or explicit goodness (e.g., good, adequate, poor). Grading is most commonly done through letter grades (e.g., A, B, C, D, E).

Relative standards. *Relative* standards express performance in comparison to others (Scriven, 1991). This is commonly done by ranking (Best, Average, Worst), but can also be done through apportioning which allocates comparative merit (including benefit, e.g., funding) across the spectrum (Top 20%, Middle 60%, Bottom 20%). A mixture of grading (an absolute determination) and ranking (a relative determination) occurs when "grading on a curve" which includes a predetermined ranking scheme. If this is the case, the grades are actually relative determinations of merit (Scriven, 1991). A flaw of relative determinations is that a poor performance could appear "good" if the rest of the group performed even more poorly (Davidson, 2005; Scriven, 1991)[4].

To ensure logical alignment, decisions about how to express standards in the rubric need to be guided by the evaluation questions. For example if the evaluation questions are in regard to the *extent* that basic needs have been met, then the standards need to be developed to directly answer that question (e.g., met, partially met, unmet). It would be inappropriate to answer an extent question in terms of *excellence.* However if the question were phrased as, *How well did the program meet participant needs*?, the answer to the question could be that the program was *excellent* at meeting participant needs. Figure 18 shows examples of parallel construction of standards to evaluation questions.

[4] A case could be made that every absolute determination has an element of comparison to it, because to know that something is *good* it does have to be compared against a standard of some sort, which is what evaluators are doing when using a rubric—outlining what *good* looks like and comparing performance against that description.

Absolute		**Relative**	
Evaluation Question	**Standards**	**Evaluation Question**	**Standards**
Is X (the evaluand) good, adequate, poor?	Grading (locate evaluand on an ordered set of categories) • Excellent (A) • Very Good (B) • Good (C) • Adequate (D) • Poor (E)	Is X (the evaluand) better than Y (a competitor)?	Ranking • Superior • Inferior
How good (or bad) is X (the evaluand)?	Scoring (interval scale) • Good (3) • Adequate (2) • Poor (1)	How much better is X (the evaluand) than Y (a competitor)?	Ranked score • 90th percentile • 50th percentile • 10th percentile
How much is X (the evaluand) worth? (How worthwhile is X?)	Appraising • Estimate of worth or value for money	Is X (the evaluand) worth more than Y (a competitor)? (More valuable or worthwhile)	Ranked appraisal • More worthwhile • Less worthwhile
How cost-effective is X?	• Not cost-effective • Minimally cost-effective • Very cost-effective	What mixture of X and Y is worth most?	Apportioning • Optimal combination of each X and Y • Semi-optimal combinations • Suboptimal combinations
To what extent is X meeting existing need?	• Needs are unmet • Partially met • Fully met	In what respects is X good?	Analytic apportioning • Optimal performance and integration of dimensions of X • Semi-optimal • Suboptimal
How effective is X?	• Not at all effective • Minimally effective • Moderately effective • Very effective	How effective is X compared to Y?	• Less effective • As effective • More effective
What is the impact of X?	• No impact • Low impact • Medium impact • High impact	What is the impact of X compared to Y?	• Lower impact • Equal impact • Higher impact

Figure 18. Examples of matching absolute and relative evaluation questions to standards (based on Michael Scriven's *Logic of Evaluation*, 1981).

Figure 19 shows the characteristics of standards.

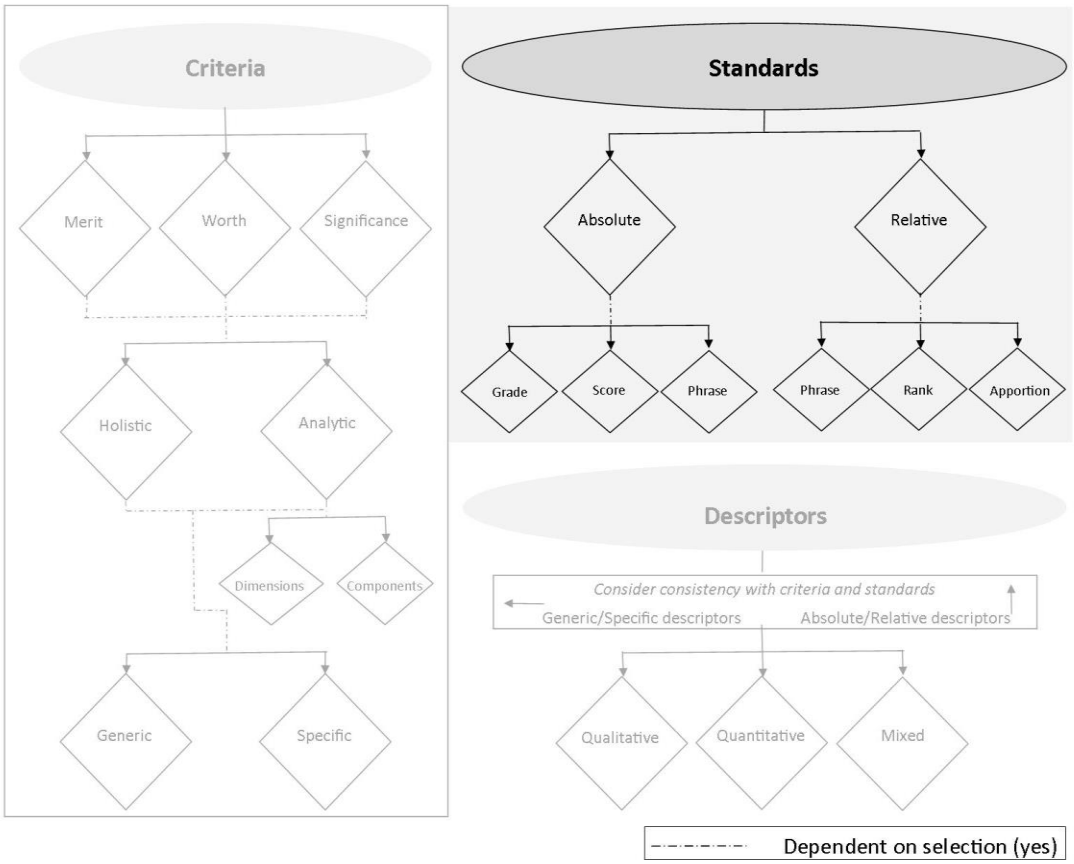


Figure 19. Flowchart of decisions when developing standards for a rubric.

Figure 20 provides examples of absolute and relative standards in rubric format.

Standards						
Absolute →	Grades	*A*	*B*	*C*	*D*	*E*
	Scores	*4*	*3*	*2*	*1*	*0*
	Extent rating	*Met Need*		*Partially Met Need*		*Unmet Need*
	Life-cycle rating	*Mature*	*Proficient*		*Developing*	*Emerging*
	Evaluative rating	*Excellent*	*Very Good*	*Good*	*Fair*	*Poor*
Relative →	Ranking: numerical	*1st*	*2nd*	*3rd*	*4th*	*5th*
	Ranking: textual	*Superior*	*Above Average*		*Average*	*Below Average*
	Apportioning*	*5%*	*20%*	*50%*	*20%*	*5%*

*E.g., A performance-based distribution scheme for funding.

Figure 20. Examples of absolute and relative standards.

Descriptors

Descriptors are the inner cells of the rubric that designate what performance should look like for each criterion across the standard spectrum. Descriptors need to convey a lot of information in a small amount of space. They can be crafted to be richly descriptive, by carefully selecting each word for succinctness, or they can be quantitatively precise (using cut scores, percentages, effect sizes), dependent on whether rubrics are synthesizing purely quantitative, purely qualitative, or a mixture of the two (Davidson, 2005).

Absolute versus relative. The wording of descriptors must align with the absolute or relative determination as set out by the standards. If *absolute* determinations are being made (grading, scoring, descriptive label), the descriptors will be written to describe the spectrum of *exemplary practice* for each criterion, whereas the text for *relative* determinations (ranking, apportioning, descriptive label) will be written to describe the spectrum of *superior practice in comparison to other programs* for each criterion (Davidson, 2005).

Specific versus generic. The choice to develop descriptors specific to the unique qualities of the program should align with criteria that were also developed to define what uniquely matters in the program. Likewise, if the criteria of merit are generic, then the descriptors can be, too. But if generic criteria of merit are selected for the rubric, such as the OECD DAC (2015) criteria contained in Figure 17, that does not constrain the descriptors to also be general. The descriptors could instead describe how the program specifically addresses exemplary or superior practice in those general categories of merit. It is also possible to develop program-specific criteria paired with generic descriptors,

although that may mean that the hard work of sense-making is left to the end of the evaluation (Martens, 2015b). Figure 21 shows characteristics of descriptors.

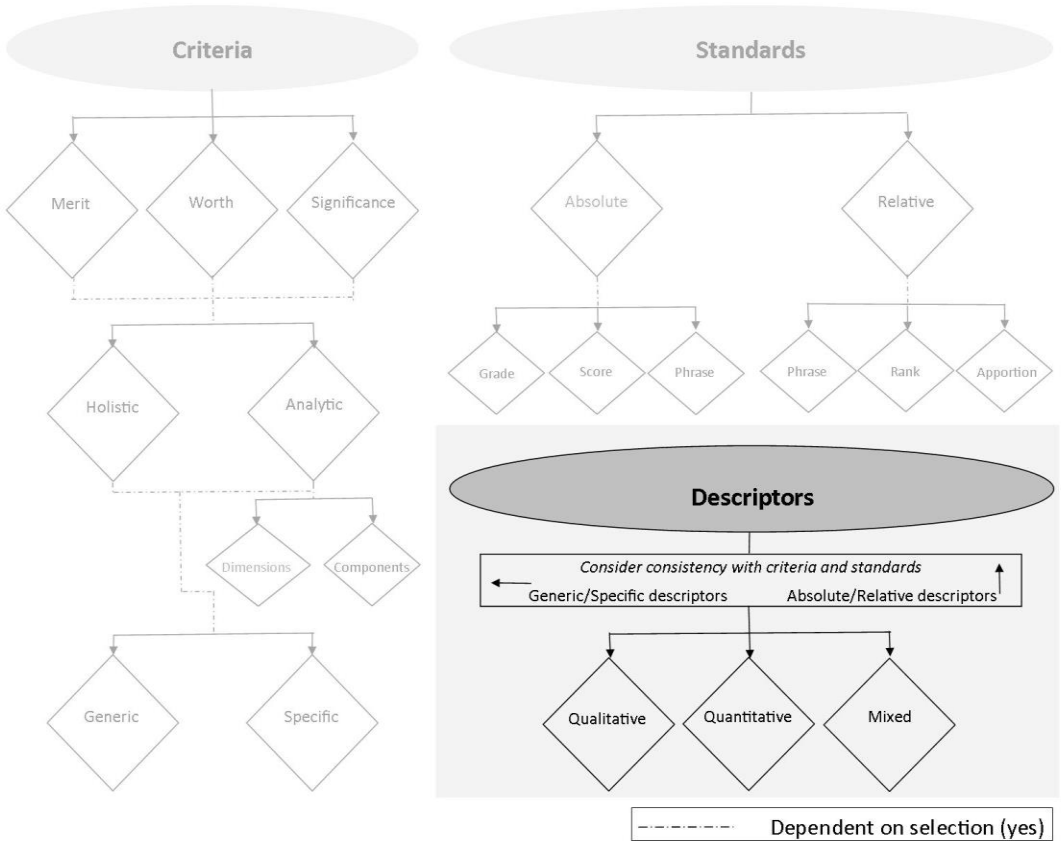


Figure 21. Flowchart of decisions when developing descriptors for a rubric.

It is best practice to incorporate a mixture of evidential aspects into a rubric to describe what the spectrum of goodness would look like, although, as the mixture of evidence builds in each cell, the ability to judge consistently is jeopardized (Davidson, 2014). To demonstrate how rubric components can be put together, Figure 22 shows an example of the flowchart in use.

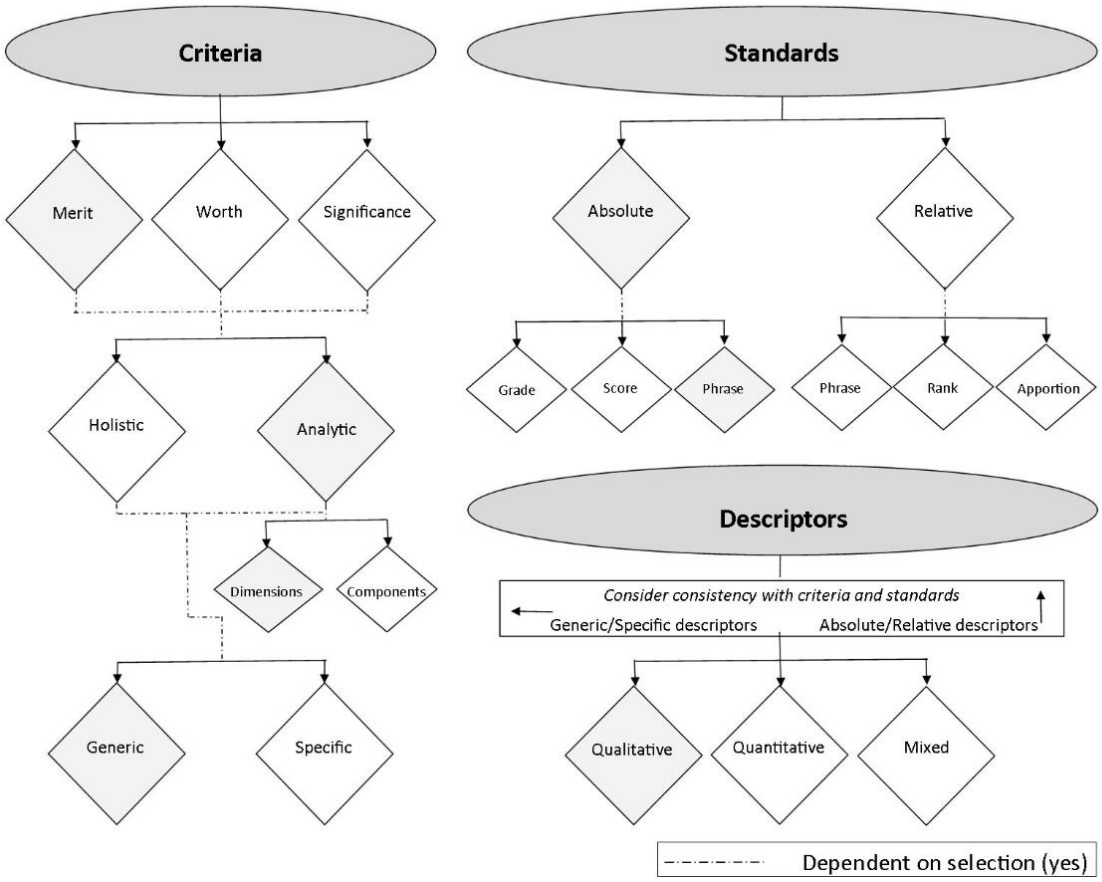


Figure 22. Flowchart of an analytic rubric with generic criteria making absolute determinations of merit via generic descriptors.

Figure 23 shows a rubric that matches the choices made in Figure 22.

Nutrition and Environmental Services:

- Students may have access to foods and beverages in a variety of venues at school including the cafeteria, vending machines, grab 'n' go kiosks, schools stores, concession stands, classroom rewards, classroom parties, school celebrations, and fundraisers.
- Premise: Healthy eating has been linked in studies to improved learning outcomes and helps ensure that students are able to reach their potential.

	Excellent	Adequate	Poor
Offerings	School nutrition services provide meals that meet federal nutrition standards for the National School Lunch and Breakfast Programs, accommodate the health and nutrition needs of all students, and help ensure that foods and beverages sold outside of the school meal programs (i.e., competitive foods) meet Smart Snacks in School nutrition standards and pricing strategies.	School nutrition services provide meals that meet federal nutrition standards for the National School Lunch and Breakfast Programs, but have a one size fits all approach, and there is little concern about the nutritional standard or pricing strategy of foods and beverages sold outside of the school meal programs (i.e., competitive foods).	School nutrition services provide meals to students, but and there is no concern about the nutritional standard or pricing strategy of foods and beverages sold outside of the school meal programs (i.e., competitive foods).
Facilitators	School nutrition professionals meet minimum education requirements and receive annual professional development and training to ensure that they have the knowledge and skills to provide these services.	School nutrition professionals meet minimum education requirements, but there are not provided professional development (PD) opportunities; OR do not meet education requirements, but have PD opportunities to upskill.	Some school nutrition staff do not meet minimum education requirements and are not provided with professional development.
Learning Opportunities	The school nutrition environment provides students with opportunities to learn about and practice healthy eating through available foods and beverages, nutrition education, and messages about food in the cafeteria and throughout the school campus.	The school nutrition environment provides students with some opportunities to learn about and practice healthy eating through available foods and beverages, nutrition education, and messages about food in the cafeteria and throughout the school campus.	The school nutrition environment gives little to no attention to opportunities to learn and practice healthy eating through available foods and beverages, nutrition education, and messages about food in the cafeteria and throughout the school campus.
Support	Most to all individuals in the school community support a healthy school nutrition environment by marketing and promoting healthier foods and beverages, encouraging participation in the school meal programs, role-modeling healthy eating behaviors, and ensuring that students have access to free drinking water throughout the school day.	Some individuals in the school community support a healthy school nutrition environment by participating in some of the following: marketing and promoting healthier foods and beverages, encouraging participation in the school meal programs, role-modeling healthy eating behaviors, and ensuring that students have access to free drinking water throughout the school day.	Few individuals in the school community support a healthy school nutrition environment by participating in a few of the following: marketing and promoting healthier foods and beverages, encouraging participation in the school meal programs, role-modeling healthy eating behaviors, and ensuring that students have access to free drinking water throughout the school day.

Figure 23. Example rubric for Figure 22 flowchart. Developed (see *Excellent* column) based on Centers for Disease Control and Prevention (2015) definitions of Components of the Whole School, Whole Community, Whole Child (WSCC) initiative.

Discussion

The aim of this study was to present how the form and function of rubrics are based on the core logic and nature of evaluation (Scriven, 1981; 1991). This objective was undertaken to fill a major gap in the program evaluation literature (Martens, 2015b). It was accomplished by positioning rubrics in evaluation logic and theory, then by describing the components of a rubric in terms of the nature and logic of evaluation and, in so doing, illustrating how the function of a rubric can make evaluative reasoning explicit and transparent.

This explanation can promote movement toward a shared language that will enable theorists, researchers, trainers, and practitioners, who often hail from disparate academic backgrounds, to more effectively further theory, training, and practice of rubric use by program evaluators to make evaluative reasoning explicit.

It is important to note that whilst a rubric can be designed to be explicitly evaluative, like any tool, it has to be *applied* (evaluatively) to *be* evaluative. If the tool is developed, but the evaluators do not use it to draw evaluative conclusions, but instead use it in non-evaluative ways, such as occurs in classification studies, then the tool is not fulfilling its evaluative purpose. It takes an evaluative rubric *and* evaluative application to make evaluative reasoning explicit.

Future research on rubrics in program evaluation should focus on informing and guiding best practices in rubric use and development for program evaluators.

References

Alkin, M. (2012). *Evaluation roots*: *A wider perspective of theorists' views and influences.* Thousand Oaks, CA: Sage.

American Evaluation Association. (2014, January 10). *What is Evaluation?* Retrieved from http://www.eval.org/p/bl/et/blogid=2&blogaid=4

Arens, S. A. (2005). *A study of evaluative reasoning in evaluative studies judged "outstanding"* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (304986745).

Australasian Evaluation Society. (2013). *Evaluators' professional learning competency framework.* Retrieved from http://aes.asn.au/professional-learning-resources.html

Cardin, F., & Alkin, M. C. (2012). Evaluation roots: An international perspective. *Journal of MultiDisciplinary Evaluation, 8*(17), 102-118.

Canadian Evaluation Society. (2010). *Competencies for Canadian evaluation practice* (V 11.0). Retrieved from http://www.evaluationcanada.ca/competencies-canadian-evaluators

Centers for Disease Control and Prevention. (2015). *Components of the whole school, whole community, whole child (WSCC) initiative*. Retrieved from http://www.cdc.gov/healthyschools/wscc/components.htm

Clements, P. (2012). Evaluating the cost effectiveness of heifer international country programs. *Journal of Multidisciplinary Evaluation, 8*(18), 16-28.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davidson, E. J. (2013). *Actionable evaluation basics: Getting succinct answers to the most important questions*. Auckland, New Zealand: Real Evaluation.

Davidson, E. J. (2014). Evaluative reasoning. *Methodological Briefs: Impact Evaluation 4*, Florence, Italy: United Nations Children's Fund (UNICEF) Office of Research.

Dirlam, D., & Byrne, M. (1978, March). *Standardized Developmental Ratings*. Paper presented at the annual 62nd meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Fournier, D. (1995). Editor's notes in reasoning in evaluation: Inferential links and leaps. *New Directions for Evaluation*, 68, 1-4.

Holistic. (2015). In *Merriam-Webster dictionary*. Retrieved from http://www.merriam-webster.com/dictionary/holistic

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), 11-20.

Martens, K. S. R. (2015a). *Rubrics in program evaluation*. Unpublished manuscript.

Martens, K. S. R. (2015b). *How program evaluators use and learn to use rubrics to make evaluative reasoning explicit.* Unpublished manuscript.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research and Evaluation, 7*(3), n.p.

OECD DAC. (2015). Criteria for Evaluating Development Assistance. Retrieved from http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm

Patton, M. Q. (2008). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership, 55*(2), 72-75.

Scriven, M. (1981). *Logic of evaluation*. Iverness, CA: Edgepress.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Scriven, M. (2000, revised 2005, 2007). The logic and methodology of checklists. Western Michigan University. Retrieved from https://www.wmich.edu/sites /default/files/attachments/u350/2014/logic%26methodology_dec07.pdf

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*(4), n.p.

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). San Francisco, CA: Jossey-Bass.

Walker, H. (2010). *Evaluation Rubric for IPod Apps*. John Hopkins University. Retrieved from: http://static.squarespace.com/static/50eca855e4b0939ae8bb12d9/50ecb58ee4b0 b16f176a9e7d/50ecb595e4b0b16f176aaab8/1288148200553/AppRubric.pdf

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

V. CONCLUSION

This three-article dissertation explores the use of rubrics as a tool for reaching explicitly evaluative conclusions in program evaluation. This chapter 1) provides an overview of the findings from the three studies; 2) discusses limitations; and 3) provides suggestions for future research.

Summary

This dissertation began with the premise that evaluation is the systematic determination of the merit, worth, significance (AEA, 2014). Throughout the manuscript, the core argument is made that it is a professional imperative for program evaluators to present transparent and explicit evaluative conclusions. However, in practice, understanding the reasoning process underlying evaluative conclusions has been problematic (Fournier, 1995). This issue is illustrated by a study conducted by Arens (2005) where she showed that the conclusions and recommendations of the studies she examined relied heavily on implicit criteria. Rubrics are tools that facilitate explicitly evaluative reasoning for evaluative conclusions and as such, rubrics should be included among the pantheon of evaluation tools and resources. This dissertation investigates rubrics and their applications in program evaluation.

Study 1 provides an account of the history and etymology of the term rubric and is followed by a review of peer-reviewed program evaluation literature. The study produced few examples of the use or described utility of rubrics to reach explicitly evaluative conclusions in program evaluation. Results from Study 2 found that practitioners who were identified as using rubrics to reach explicitly evaluation

conclusions actually find them to be useful in multiple ways (including reaching explicitly evaluative conclusions), but they are rarely publishing their experiences in the peer-reviewed literature. Guidance is instead typically shared through informal mentoring. Study 3 describes how the form and function of rubrics assist program evaluators to reach explicitly evaluative conclusions.

Bringing the three studies together, it was shown in Study 1 that the term rubric has a long history of being used to draw attention to important detail, such as the labeling of sheep, the cut-line on wood, and significant text. Evaluators who use rubrics as evaluation-specific methodologies continue this tradition, because they draw out the high-level, important dimensions of a program, outline the standards of merit as levels of performance by which the dimensions will be judged through descriptors that are developed to exemplify what performance would look like for each dimension at each level of performance. So, a rubric explicitly draws attention to the reasoning in drawing evaluative conclusions.

Study 1 also points out that evaluators hail from diverse academic backgrounds, which Study 2 supports. Diversity in background and thought can enhance creative functioning, but lack of homogeneity, especially when it comes to shared language, can prove to be problematic and create misunderstandings if not addressed. Study 3 provides a clear description of the components of a rubric using the language of the logic and nature of evaluation. This may serve to provide evaluation practitioners with a nomenclature that can enhance communication and provide a stronger foundation on which "shared language" can be developed. As a further extension of my research, in Figure 24, I present the criteria of merit for a rubric. This list brings together information

from evaluation theorists, the peer-reviewed program evaluation literature, parallel literature, and interviews with program evaluators who use rubrics as an evaluation-specific methodology. The list can assist practitioners in development or selection of sound rubrics.

Rubric Criteria of Merit

- Structure
  - Criteria—from the Logic and Methodology of Checklists (Scriven, 2000, revised 2005, 2007)
    - Criteria (not indicators)
    - Complete (no significant omissions)
    - Contiguous (non-overlapping)
    - Commensurate (of the same level—which indicates *important aspects* of the evaluand)
    - Clear (comprehensible, applicable)
    - Concise (no superfluous criteria—one to five if possible)
    - Confirmable (measurable or reliably inferable)
  - Standards
    - Logical spectrum
    - Enough levels to be discrete, yet discriminate
    - Spectrum of parallel construction to answer/inform evaluation question(s)
    - Spectrum uses evaluative labels (grading, scoring, ranking, apportioning, or phrasing)
  - Descriptors
    - Differentiating
    - Descriptive, yet concise
    - Clear, understandable, and free from jargon
- Purpose
  - Content/focus is clear
  - Fit for task
  - Appropriate level of specificity
  - Credible (appropriate level of refinement/measurement testing)
- Development
  - As appropriate, based on a mixture of:
    - Need
    - Literature
    - Research
    - Stakeholder input
  - Purposefully builds in flexibility through an appropriate mix of:
    - Expectations for revision
    - Catchall (slush) criterion
    - Scheduled revision checks
- Presentation
  - Formatted to enhance understanding
  - Visually accessible (e.g., displays on one page)
  - Explicitly conveys evaluative reasoning
  - Promotes discussion/negotiation of shared understanding and sense making efforts
  - In sum, the rubric elements comprise a comprehensive and plausible depiction of the spectrum of the evaluand's quality

Figure 24. Criteria of merit for a rubric.

Limitations

The review of program evaluation literature that comprised Study 1 spanned close to 40 years and initially included 239 articles, but only resulted in 20 examples of rubric use or described utility in program evaluation. This suggests that the use of rubrics in program evaluation is somewhat rare. Although, there is the possibility that rubric use in program evaluation is actually more prevalent in practice, but is just not a topic that makes it into the peer reviewed literature. If this is the case, then this research fills a critical gap in what we know about the use of rubrics in program evaluation.

Due to the inclusion criterion for articles in Study 1 (peer-reviewed publications about rubric use or utility in program evaluation) and interviewees in Study 2 (individuals with relevant experience using rubrics as an evaluation-specific methodology), the population for Study 1 and the potential population for Study 2 were small. The initial sample for Study 2 included 11 individuals identified as having published (peer-reviewed literature, and evaluation books) or engaged in dissemination on the topic (e.g., conference presentations). To reach more potential participants, Study 2 was designed with a snowball sampling method. Employing this method was also an effort to provide unpublished evaluation practitioners the opportunity to share their experience and guidance on rubric use as an evaluation-specific methodology. The pool broadened into 18 possible interviewees, 10 of whom participated. Of the ten interviews, only nine interviews were retained as meeting the criteria of interviewees using rubrics as evaluation-specific tools. While the small numbers in both studies (20 articles and 9 interviews) might be seen as a limitation, but they were reflective of their respective

universes. Also, because the response rate in Study 2 was reasonable, the findings appear to be generalizable.

Suggestions for Future Research

Future research should focus on investigating proven practices for developing and using rubrics in program evaluation. It is possible to continue to harvest practices from parallel literature to inform rubric development and use in program evaluation. It is also possible to cast a wider net in an attempt to include more evaluation practitioners who undertake this practice via invitation postings on evaluation electronic mailing lists. Electronic mailing lists could be hosted by such organizations as the: American Evaluation Association (AEA); Canadian Evaluation Society (CES); Aotearoa New Zealand Evaluation Association (ANZEA); Australasian Evaluation Society (AES); and European Evaluation Society (EES).

Other avenues for future research concern investigating how rubric use affects evaluation quality (perhaps through an investigation of evaluation reports based on rubric use) and how rubric use affects evaluation use (perhaps through investigating the acceptance, influence, and/or impact of evaluations that included rubrics as evaluation-specific methodologies). These types of studies will inform the value-added benefit of the use of rubrics. During interviews conducted for Study 2 of this dissertation, participants indicated that funders are beginning to ask for, or even provide, rubrics in their calls for proposals. One of the individuals who relayed this information was the interviewee who was excluded from the findings because s/he was honest about not using rubrics to draw evaluative conclusions. But this information may be important, as it may indicate that

funders are hungry for the practice to expand. Yet, not enough program evaluators appear to be practicing rubric use to effectively respond to this possible wider call. Thus future research can form a foundation for development of future guidance, training, and resources that support evaluators in expanding their evaluation-practice repertoire.

References

American Evaluation Association. (2014, January 10). *What is Evaluation?* Retrieved from http://www.eval.org/p/bl/et/blogid=2&blogaid=4

Arens, S. A. (2005). *A study of evaluative reasoning in evaluative studies judged "outstanding"* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (304986745).

Fournier, D. (1995). Editor's notes in reasoning in evaluation: Inferential links and leaps. *New Directions for Evaluation,* 68, 1-4.

Scriven, M. (2000, revised 2005, 2007). The logic and methodology of checklists. Western Michigan University. Retrieved from https://www.wmich.edu/sites /default/files/attachments/u350/2014/logic%26methodology_dec07.pdf

APPENDIX A

Human Subjects Institutional Review Board Letter

Date: September 3, 2014

To: Lori Wingate, Principal Investigator
Krystin Martens, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed for HSIRB Project Number 14-09-06

This letter will serve as confirmation that your project titled "Rubrics as a Tool for Reaching Explicitly Evaluative Conclusions: Implications for Evaluation Theory, Training, and Practice" has been reviewed by the Human Subjects Institutional Review Board (HSIRB). Based on that review, the HSIRB has determined that approval is not required for you to conduct this project because you are analyzing the development and use of rubrics and not collecting personal identifiable (private) information about an individual.

Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.

APPENDIX B

Interview Participant Email Invitation

Dear [Name]

Your work involving rubric use in program evaluation has come to my attention. I am inviting academics and practitioners identified through my review of peer-reviewed and grey literature, and individuals identified by the initial interviewees, to participate in a research project titled, "How Evaluators Learn the Art and Science of Rubric Development and Use," which is one of three stand-alone studies that when combined will comprise a larger research project titled "Rubrics as a Tool for Reaching Explicitly Evaluative Conclusions: Implications for Evaluation Theory, Training, and Practice." This larger project will serve as my dissertation research project for the requirements of a Ph.D. in Interdisciplinary Evaluation at Western Michigan University. My committee consists of Lori Wingate, Michael Scriven, and Jane Davidson. More information is provided in the attached document.

To inform the study, I am conducting 1 hour interviews with academics and practitioners like yourself. The interviews conducted thus far have lasted about 45 minutes, but I am asking to schedule one hour to ensure enough time. In the interview, you will be asked to share things such as how you were trained to develop and use rubrics and how you use them in your evaluation practice.

Please review the attached information about the study and reply back to this email if you are interested in participating. We can then arrange a time to meet (in person, by phone, or by Skype or Zoom).

I look forward to hearing from you.

Krystin Martens

*This study has been reviewed by the Human Subjects Institutional Review Board (HSIRB) at Western Michigan University. The board determined that approval is not required to conduct this project because the analysis focuses on the development and use of rubrics and we are not collecting personal identifiable (private) information about an individual.

APPENDIX C

Semi-Structured Interview Protocol

Hello. I am Krystin Martens. Thank you so much for participating in my dissertation research. I want to remind you that the interview will last about 60 minutes. You do not have to answer any question you don't feel like answering and you may stop the interview at any time. If it is OK with you, I would like to record our conversation. Is that OK with you?

BACKGROUND AND USE

1. Please start out by telling me a bit about your background in evaluation.
   - Training in evaluation (formal, informal)
   - Years practicing evaluation
   - Background discipline context (if any)
   - Practice discipline context (if any)
   - What is the highest degree you have attained?
   - How long have you been using rubrics in program evaluation?
   - In what professional evaluation associations or societies are you a member?
2. How do you use rubrics in your practice?
3. What makes for a good rubric / what are the essential elements of a useful/practical rubric?
4. What makes for a good rubric development process? How/when do you involve stakeholders?
5. Do you engage in testing and refining rubrics? If so, can you describe that process
6. Are there certain contexts that you have found that rubrics work better for you than others? (Evaluation approaches; disciplines; stakeholder make-ups; audience make-ups)
7. What are the downsides/challenges to rubric use in evaluation?
8. Under what circumstances should you NOT use rubrics?

TRAINING

9. Where and how did you learn about using and developing rubrics?
10. What advice do you have for others interested in learning how to use rubrics in their practice?
11. Do you train other practitioners in rubric development and use? If so, how?

IMPORTANCE

12. How important do you believe training in the development and use of rubrics is for evaluation practitioners?
13. What other ways have you used to explicitly derive evaluative conclusions? Defined as: Conclusions that use value terms to denote the goodness of what is being evaluated, and clearly communicate the logic and reasoning used to come to those conclusions.
14. How often do you use rubrics in your evaluations? (% of time)
15. If you never started using rubrics, what would your evaluation practice be like?

CONCLUSION

16. Is there anything that I should have asked you, but didn't?
17. Or anything else that you would like to add?

SNOWBALL SAMPLE

18. Do you know of others who develop or use rubrics in evaluation whom I might contact for the study? Do you have contact information for that person? Is it OK if I mention to him/her that you recommended him/her as a study participant?

Thank you for taking the time to speak with me. I will write up my notes and send them to you for verification by _____[date].

APPENDIX D

Figure Reproduction Release Example

[Date]

Dear [Name]:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

[Citation]

[Figure image]

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[contact information]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

_____

[name]

Date:_____

APPENDIX E

Letters of Permissions

February 25, 2016

Dear Chris Coryn:

I am contacting you because you are the Executive Editor of the Journal of Multidisciplinary Evaluation.

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), p. 14; Table 1: *Example of a generic rubric*.

Table 1
Example of a Generic Rubric

	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Chris Coryn, Executive Editor of the Journal of Multidisciplinary Evaluation

Date: February 26, 2016

February 5, 2016

Dear Julian King:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), p. 14; Table 1: *Example of a generic rubric.*

Table 1
Example of a Generic Rubric

	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in

the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Julian King

Date: 19 February 2016

February 5, 2016

Dear Kate McKegg:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), p. 14; Table 1: *Example of a generic rubric.*

Table 1
Example of a Generic Rubric

	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in

the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Kate McKegg

Date: 19/2/16

February 5, 2016

Dear Judy Oakden:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), p. 14; Table 1: *Example of a generic rubric*.

Table 1
Example of a Generic Rubric

	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in

the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

J P Oakden

Judy Oakden

Date: 25 February 2016

February 5, 2016

Dear Nan Wehipeihana:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation, 9*(21), p. 14; Table 1: *Example of a generic rubric*.

Table 1
Example of a Generic Rubric

	Criteria for Rating Answers to Key Evaluation Questions
Excellent	Performance is clearly very strong or exemplary in relation to the question. Any gaps or weaknesses are not significant and are managed effectively.
Good	Performance is generally strong in relation to the question. No significant gaps or weaknesses, and less significant gaps or weaknesses are mostly managed effectively.
Adequate	Performance is inconsistent in relation to the question. Some gaps or weaknesses. Meets minimum expectations/requirements as far as can be determined.
Poor	Performance is unacceptably weak in relation to the question. Does not meet minimum expectations/requirements.
Insufficient Evidence	Evidence unavailable or of insufficient quality to determine performance.

Note: Adapted from NZQA (2009).

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in

the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

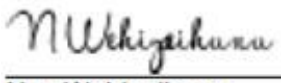
Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:



Nan Wehipeihana

Date: 19/02/2016

February 25, 2016

Dear William Wallace,

I am contacting you because you are the Executive Officer of the Australasian Evaluation Society, which is the publisher of the Evaluation Journal of Australasia.

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

Dickinson, P. & Adams, J. (2012). Building Evaluation Capability in the Public Health Workforce: Are Evaluation Training Workshops Effective and What Else Is Needed? *Evaluation Journal of Australasia, 12*(2), p. 28-39; Table 1: *Merit determination for quality—Three-day and one-day workshops*, p. 32.

TABLE 1: MERIT DETERMINATION FOR QUALITY—THREE-DAY AND ONE-DAY WORKSHOPS

Excellent	Nearly all participants (95%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Nearly all would recommend the workshops to others and no problems were identified regarding the quality of workshop content, design or delivery.
Very Good	Most participants (75–94%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Most would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Good	At least half of participants (50–74%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. At least half would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Poor	Less than half (<50%) of participants either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Less than half would recommend the workshops to others OR major problems were identified regarding the quality of workshop content, design or delivery and these were not managed effectively.

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

William Wallace,
Executive Officer of the Australasian Evaluation Society: publisher of the Evaluation Journal of Australasia

Date: 25 February 2016

Reply Reply All Forward

RE: Permission to reproduce figure

Dickinson, Pauline [P.M.Dickinson@massey.ac.nz]

To: Krystin Sue Martens

Saturday, 24 October 2015 2:21 PM

- You replied on 25/10/2015 7:48 AM.

From: Dickinson, Pauline [P.M.Dickinson@massey.ac.nz]
Sent: Wednesday, 21 October 2015 2:30 PM
To: Krystin Sue Martens; Adams, Jeff
Subject: RE: Permission to reproduce figure

Hi krystin

That is fine to use the table

Regards

Pauline

From: Krystin Sue Martens [krystin.martens@unimelb.edu.au]
Sent: Wednesday, October 21, 2015 4:26 PM
To: Adams, Jeff; Dickinson, Pauline
Subject: Permission to reproduce figure

Hi Dr. Adams & Dr. Dickinson,

Your ECB articles in AJE (2010) and EJA (2012) are included in a review I have written for publication on rubric use in program evaluation. The article is part of a three article dissertation for a Ph.D. in evaluation from Western Michigan University.

I would like to include examples found in my review to show how rubrics are being used and would like your permission to reproduce the below table from:

Dickinson, P., & Adams, J. (2012). Building evaluation capability in the public health workforce: Are evaluation training workshops effective and what else is needed? Evaluation Journal of Australasia, 12(2), 28-39.

The excerpt is:
[X]
I am currently looking into the release requirements (forms) I will need to file to publish. I am planning to submit the article to AJE, (my second tier choices are JMDE or EJA). I also need to obtain permission for the overall dissertation manuscript that will be published through Proquest. Please let me know if you would grant me permission to reproduce your rubric.

I look forward to hearing from you.

Krystin

February 5, 2016

Dear Jeffrey Adams:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

Dickinson, P. & Adams, J. (2012). Building Evaluation Capability in the Public Health Workforce: Are Evaluation Training Workshops Effective and What Else Is Needed? *Evaluation Journal of Australasia, 12*(2), p. 28-39; Table 1: *Merit determination for quality—Three-day and one-day workshops*, p. 32.

TABLE 1: MERIT DETERMINATION FOR QUALITY—THREE-DAY AND ONE-DAY WORKSHOPS

Excellent	Nearly all participants (95%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Nearly all would recommend the workshops to others and no problems were identified regarding the quality of workshop content, design or delivery.
Very Good	Most participants (75–94%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Most would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Good	At least half of participants (50–74%) either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. At least half would recommend the workshops to others. Any problems regarding the quality of workshop content, design or delivery were managed effectively.
Poor	Less than half (<50%) of participants either agreed or strongly agreed that: the workshop content was relevant; they had time to practise skills; and the presentations held their attention. Less than half would recommend the workshops to others OR major problems were identified regarding the quality of workshop content, design or delivery and these were not managed effectively.

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to

supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Jeffrey Adams

Jeffery

Date: 10/2/16

Copyright Clearance Center RightsLink®

My Orders | My Library | My Profile | Welcome krystin.martens@unimelb.edu.au | Log out | Help

My Orders > Orders > All Orders

License Details

Thank you very much for your order.

This is a License Agreement between Krystin S Martens ("You") and Elsevier ("Elsevier"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

Get the printable license.

License Number	3816311114676
License date	Feb 26, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Evaluation and Program Planning
Licensed Content Title	A framework to assess the development and capacity of non-profit agencies
Licensed Content Author	Russell G. Schuh,Laura C. Leviton
Licensed Content Date	May 2006
Licensed content volume number	29
Licensed content issue number	2
Number of pages	9
Type of Use	reuse in a journal/magazine
Requestor type	author of new work
Intended publisher of new work	Adis
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 1.
Title of the article	A framework to assess the development and capacity of non-profit agencies
Publication new article is in	Rubrics in program evaluation
Publisher of the new article	Adis
Author of new article	Krystin S. R. Martens
Expected publication date	May 2016
Estimated size of new article (number of pages)	32
Elsevier VAT number	GB 494 6272 12
Price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	**0.00 USD**

BACK

February 5, 2016

Dear Russell G. Schuh:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

Schuh, R. G. & Leviton, L. C. (2006). A framework to assess the development and capacity of non-profit agencies. *Evaluation and Program Planning, 29*(2), p. 171-179; Figure 1. *Example of scoring rubric for financial resources* p. 175.

Features		Agency Stage of Development				
		Stage 1 Initial, start-up or small agency	Stage 2 Advanced young or small agency	Stage 3 Established agency	Stage 4 Proficient agency	Stage 5 Highly productive agency
Financial Resources Maturity (sample attributes)	*A*	Unstable, unpredictable or insecure funding	Limited funding, more stable than stage 1 agency	Established agency funding where organization has some control over its own destiny	Stable sources of income or funding	Highly stable funding
	B	Opportunistic funding (like that resulting from a new or emerging policy concern).	Contingency funding for activities or services but not long term (e.g. summer youth activities).	Initiative funding - Project funding derived from services related to an on-going policy priority like violence prevention or early childhood development.	Niche funding - Agency funded through on-going source like United Way or local funding authority (e.g. mental health, housing or juvenile justice).	Institutionalized funding for at least a core of services or programs (e.g. Funding is a line item on a state or local authority budget.
	C	Start-up or seed money only or short term funding.	Projects or activities funded on a one-time basis (funding may span several years).	Funding may be for specified periods but contains provision for renewal or continuation.	Core funding on-going, may have provision for periodic review or renewal.	Core service funding is highly stable. May be supported by spin-off services or products.

Fig. 1. Examples of scoring rubric for financial resources.

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to

supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Russell G Schuh

Russell G. Schuh
schuh@pitt.edu

Date: 02/24/2016

February 5, 2016

Dear Laura Leviton:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to include an excerpt from the following item in my dissertation:

Schuh, R. G. & Leviton, L. C. (2006). A framework to assess the development and capacity of non-profit agencies. *Evaluation and Program Planning, 29*(2), p. 171-179; Figure 1. *Example of scoring rubric for financial resources* p. 175.

Features		Agency Stage of Development				
		Stage 1 Initial, start-up or small agency	Stage 2 Advanced young or small agency	Stage 3 Established agency	Stage 4 Proficient agency	Stage 5 Highly productive agency
Financial Resources Maturity (sample attributes)	A	Unstable, unpredictable or insecure funding	Limited funding, more stable than stage 1 agency	Established agency funding where organization has some control over its own destiny	Stable sources of income or funding	Highly stable funding
	B	Opportunistic funding (like that resulting from a new or emerging policy concern).	Contingency funding for activities or services but not long term (e.g. summer youth activities).	Initiative funding - Project funding derived from services related to an on-going policy priority like violence prevention or early childhood development.	Niche funding - Agency funded through on-going source like United Way or local funding authority (e.g. mental health, housing or juvenile justice).	Institutionalized funding for at least a core of services or programs (e.g. Funding is a line item on a state or local authority budget.
	C	Start-up or seed money only or short term funding.	Projects or activities funded on a one-time basis (funding may span several years).	Funding may be for specified periods but contains provision for renewal or continuation.	Core funding on-going, may have provision for periodic review or renewal.	Core service funding is highly stable. May be supported by spin-off services or products.

Fig. 1. Examples of scoring rubric for financial resources.

The figure would appear in Study 1 of my dissertation entitled: "Rubrics in Program Evaluation." Study 1 is review of peer-reviewed evaluation literature. In this study, I selected examples from the literature to illustrate rubric use in program evaluation. The source will receive full credit in the manuscript and in the article.

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to

supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Laura Leviton

[redacted]

Date: 2/19/2016

February 22, 2016

Dear Harry Walker:

I am completing a doctoral dissertation at Western Michigan University entitled "Rubrics as a tool for reaching explicitly evaluative conclusions: Implications for evaluation theory, training, and practice." My dissertation is composed of three publishable articles. I would like to request your permission to adapt the following item in my dissertation:

Evaluation Rubric for IPod Apps

Domain	1	2	3	4
Curriculum Connection	Skill(s) reinforced in the app are not clearly connected to the targeted skill or concept	Skill(s) reinforced are prerequisite or foundation skills for the targeted skill or concept	Skill(s) reinforced are related to the targeted skill or concept	Skill(s) reinforced are strongly connected to the targeted skill or concept
Authenticity	Skills are practiced in a rote or isolated fashion (e.g., flashcards)	Skills are practiced in a contrived game/simulation format	Some aspects of the app are presented an authentic learning environment	Targeted skills are practiced in an authentic format/problem-based learning environment
Feedback	Feedback is limited to correctness of student responses	Feedback is limited to correctness of student responses and may allow for student to try again	Feedback is specific and results in improved student performance (may include tutorial aids)	Feedback is specific and results in improved student performance: Data is available electronically to student and teacher
Differentiation	App offers no flexibility (settings cannot be altered)	App offers limited flexibility (e.g., few levels such as easy, medium, hard)	App offers more than one degree of flexibility to adjust settings to meet student needs	App offers complete flexibility to alter settings to meet student needs
User Friendliness	Students need constant teacher supervision in order to use the app	Students need to have the teacher review how to the use the app on more than one occasion	Students need to have the teacher review how to the use the app	Students can launch and navigate within the app independently
Student Motivation	Students avoid the use of the app or complain when the app is assigned by the teacher	Students view the app as "more schoolwork" and may be off-task when directed by the teacher to use the app	Students will use the app as directed by the teacher	Students are highly motivated to use the app and select it as their first choice from a selection of related choices of apps

Created by Harry Walker – Johns Hopkins University 10/18/2010
Please contact for permission to use hwalker@bcps.org

I would like to adapt your figure to be presented as follows:

Evaluation Rubric for IPOD APPS

	4	3	2	1
Curriculum Connection	Skills reinforced are strongly connected to the targeted skill or concept	Skills reinforced are related to the targeted skill or concept	Skills reinforced are prerequisite or foundational skills for the targeted skill or concept	Skills reinforced in the app are not clearly connected to the targeted skill or concept
Authenticity	Targeted skills are practiced in an authentic format/ problem-based learning environment	Some aspects of the app are presented in an authentic learning environment	Skills are practiced in a contrived game/simulation format	Skills are practiced in a rote or isolated fashion (e.g., flashcards)
Feedback	Feedback is specific and results in improved student performance; Data is available electronically to the student and teacher	Feedback specific and results in improved student performance (may include tutorial aids)	Feedback is limited to the correctness of student responses and may allow students to try again	Feedback is limited to the correctness of student responses
Differentiation	App offers complete flexibility to alter settings to meet student needs	App offers more than one degree of flexibility to adjust settings to meet student needs	App offers limited flexibility (e.g., few levels, such as: easy, medium, hard)	App offers no flexibility (settings cannot be altered)
User Friendliness	Students can launch and navigate within the app independently	Students need to have the teacher review how to use the app	Students need teacher to review how to use the app on more than one occasion	Students need constant teacher supervision in order to use the app
Student Motivation	Students are highly motivated to use app and select it as their choice among related choices of apps	Students will use the app as directed by the teacher	Students view app as "more school work" and may be off-task when directed by the teacher to use the app	Students avoid use of the app or complain when use of the app is assigned by the teacher

The figure would appear in Study 3 of my dissertation entitled: "How the Form and Function of Rubrics Make Evaluative Reasoning Explicit." In this study, I explicate how the three components of a rubric (criteria, standards, and descriptors) make them explicitly evaluative. Your figure, if you grant me permission to use it, would be used as an example of a rubric.

The source will receive full credit in the dissertation and in the article, as such, I would appreciate guidance on how to do so. I am thinking it may be something like:

Figure 2. Example rubric. Adapted from Walker, H. (2010). *Evaluation Rubric for IPod Apps*. John Hopkins University. Used by permission. Retrieved from: http://static.squarespace.com/static/50eca855e4b0939ae8bb12d9/50ecb58ee4b0b16f176a9e7d/50ecb595e4b0b16f176aaab8/1288148200553/AppRubric.pdf

By agreeing to the use of the item in my dissertation, you give ProQuest Information and Learning (PQIL) the right to supply copies of this material on demand as part of my doctoral dissertation. You also give the journal in which my article is ultimately published the right to supply copies of this material on demand as part of the volume in which the article is published. Please attach any other terms and conditions for the proposed use of this item.

If you no longer hold the copyright to this work, please indicate to whom I should direct my request.

Thank you for your time and attention to this matter.

Sincerely,

Krystin Martens

[redacted]

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

Harry Walker

[redacted]

Date: March 21, 2016