Harnessing Complexity: Analysis Methodology and Ethical Framework to Facilitate Utilization of Video Data in Evaluations

Kurt A. Wilson
Western Michigan University, kurt@effectx.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Policy Design, Analysis, and Evaluation Commons, and the Statistical Methodology Commons

Recommended Citation
https://scholarworks.wmich.edu/dissertations/268
HARNESSING COMPLEXITY: ANALYSIS METHODOLOGY AND ETHICAL FRAMEWORK TO FACILITATE UTILIZATION OF VIDEO DATA IN EVALUATIONS

by

Kurt A. Wilson

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 2014

Doctoral Committee:

Chris Coryn, Ph.D., Chair
Daniela Schroeter, Ph.D.
Patricia Reeves, Ph.D.
Tarek Azzam, Ph.D.
Most evaluations in the nonprofit and international development sectors are conducted in contexts of complexity; the specific intervention being evaluated is but one of many interrelated factors influencing the desired outcome. Video data, especially when directly generated by program participants, can provide both exceptionally rich qualitative data as well as contextually-relevant feedback within complex systems. Despite these unique strengths and opportunities, video data is underutilized in the field of evaluation. This dissertation addresses specific barriers associated with video data through three inter-related papers: Papers one and two (Chapters II and III) present the findings from two interrelated studies of an analysis methodology utilizing crowd-sourcing, and paper three (Chapter IV) presents a practical ethical framework for the use of visual data.

The video data analyzed in papers one and two was user-generated video reviews of various products, and both studies explored the time, cost and quality implications of crowd-sourcing video analysis. The broad conclusions presented in Paper one indicate that while there are specific limitations, the crowdsourcing methodology offers considerable potential advantages for evaluators - not just saving time and money, but also enhancing the richness of analysis through the analytic
perspective of a broad group of participants. Similarly, the findings of Paper two indicate that categorical coding of video data can be completed in a matter of hours at a reasonable cost, and that considerable accuracy is possible – though within specific parameters.

Paper three offers the conceptual framework of ‘relational integrity’ identified as a theme in the literature of ethical visual methodologies. Key themes from a comparative analysis of the image use procedures in the fields of entertainment, journalism, advertising, and social media are presented as contextual reference for guiding ethical decisions and informing relationships surrounding visual methodologies. Five categories of relationship are suggested based on this analysis to serve as a practical reference for evaluators.

While additional study is needed to extend and refine the findings from each of these studies, the findings combine to offer both new direction and practical guidance for the evaluation community working in complex contexts.
ACKNOWLEDGMENTS

The completion of this dissertation is a significant milestone on my journey with evaluation, and as I reflect on the steps that brought me to this point, I feel profound gratitude for the extended community of support that has facilitated each step.

Chris Coryn addressed my initial hesitations about being a nontraditional student in the IDPE program with inspiration and thoughtfulness, and provided consistent encouragement and valuable opportunities throughout my studies. His position as committee chair was a natural capstone to this role, and his effective leadership maintained the momentum of the project, facilitated the contributions of the rest of the committee, and helped me better understand the structure and flow of academic writing. I am deeply grateful for his formative contributions to this work, my overall understanding of evaluation, and initial steps into evaluation consulting.

Daniela Schroter provided a valuable counter-balance to my big-picture creative tendencies, and regularly brought my focus back to the important details and clear, specific writing. Her careful proofing and thoughtful notes contributed significantly to both the final shape of this project and my overall learning through the process, and I greatly appreciate the considerable time she gave toward this end.

Patricia Reeves provided both my initial introduction to qualitative methods and inspiration for the study on the ethical use of visual data in Chapter IV. I owe much of my understanding and ongoing use of qualitative methods to her influence,
Acknowledgments-continued

and I am thankful for both her encouragement of my enthusiasm and challenge to dig deeper.

Tarek Azzam co-led a workshop at the European Evaluation Society which both focused and further fueled my interest in the application of new technologies to evaluation. His enthusiasm and encouragement provided the inspiration for Chapters II and III, and his insightful guidance as I conducted the studies and wrote the papers provided significant help in sharpening both my thinking and communication. I am thankful for his thoughtful involvement and willingness to sacrifice his schedule and work across the timezones.

My approach to evaluation flows directly from my parents; I regularly feel aspects of both the analytical and rigorous approach Dr. Calvin Wilson brings to medicine and creativity and attention to communication Mimi Wilson brings to her books and speaking. I am profoundly grateful for their love, nurture and encouragement which has taken deeper and richer character with each passing year, as well as the very tangible help with a car to help with the regular commute to Kalamazoo.

I would not have started, much less finished, without Lori’s encouragement and support as my wife, thought partner and soul-mate. Her faithful help keeping the household running gave me the margin to study, and our many discussions while walking through London provided significant inspiration through the long process of
Acknowledgments-continued

the literature review and initial framing of the concepts. Pierce and Miriam provided breaks that were (usually!) welcome and always helped me engage the broader fullness of life, and their accommodation of my tight schedule and extended focus on studies is much appreciated. I’m also indebted to my in-laws, Rick & Sylvia Allen, and to the extensive constellation of family and friends who provided the affirmation and support that brought joy to the journey.

Finally, I am very grateful for the warm welcome, encouragement and challenge provided by the extended community around the Evaluation Center and IDPE program. Mary Ramlow’s patient explanation of WMU procedures and help with the forms and key details that would escape me kept me on track. Her meeting with me after my initial discussion with Chris clarified the process and solidified my decision to join the program, and I am thankful for her hard work and pivotal role in the program. Even though it involved a long commute, the opportunity to join practicing evaluators and other students working to internalize the concepts made it all worth it, and I am grateful for the opportunity to be part of such a vibrant community. As relates to this dissertation, Kelly Robertson and Jan Fields provided wonderful and timely help as part of the control group for Paper two; thank you much!
Acknowledgments-continued

The milestone marked by this dissertation is the completion of the IDPE program, my new ‘alma mater,’ or ‘fostering mother’ of a whole new direction and chapter of life.

Kurt A. Wilson
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... ii

LIST OF TABLES .................................................................................................................... xi

I. INTRODUCTION .................................................................................................................. 1

   Context of Complexity ....................................................................................................... 2

   Benefits of Visual Methods ............................................................................................... 5

   Benefits of Video Data ...................................................................................................... 8

   Problem ............................................................................................................................. 10

   Literature Review ............................................................................................................. 13

      Overview ......................................................................................................................... 13

      Visual Methods in the Social Sciences ................................................................. 15

      Underutilization of Visual Methods in Evaluation ..................................... 16

      Underutilization of Visual Methods in the Social Sciences ........... 19

   Paper 1 Summary ............................................................................................................ 22

   Paper 2 Summary ............................................................................................................ 24

   Paper 3 Summary ............................................................................................................ 26

II. CROWD-SOURCED OPEN CODING OF USER-GENERATED EVALUATION DATA ........................................... 28

   Authors .............................................................................................................................. 28

   Abstract ............................................................................................................................. 28

   Introduction ....................................................................................................................... 29
Table of Contents - continued

CHAPTER

User-Generated Video Data ......................................................... 63

Video Data Analysis ................................................................. 64

Mturk Crowd-Source Platform .................................................... 65

The Study .................................................................................. 70

Method ....................................................................................... 71

Design ......................................................................................... 71

Instrumentation ......................................................................... 72

Sample ......................................................................................... 74

Procedure ..................................................................................... 78

Analytic Approach ....................................................................... 79

Findings ....................................................................................... 81

Cohort Analysis ........................................................................... 81

Video Analysis ............................................................................ 85

Analysis of Bias ........................................................................... 87

Discussion .................................................................................. 88

Potential Benefits ......................................................................... 90

Limitations and Future Research ............................................... 91

Applications of Method .............................................................. 92

IV. RELATIONAL INTEGRITY: A CONCEPTUAL FRAME FOR ETHICAL VISUAL METHODOLOGIES INFORMED BY THE FIELDS OF ENTERTAINMENT, JOURNALISM, ADVERTISING AND SOCIAL MEDIA ............................................................... 95

Authors ....................................................................................... 95
Table of Contents - continued

CHAPTER

Abstract .......................................................................................................................... 95

Introduction .................................................................................................................. 96

Context for Visual Methods ......................................................................................... 97

History and Culture ..................................................................................................... 97

Primary Domains of Ethical Visual Methodology ..................................................... 99

Relational Integrity ...................................................................................................... 102

Focal Research Questions .......................................................................................... 110

Method .......................................................................................................................... 110

Design and Analytic Approach .................................................................................... 111

Data Sources ................................................................................................................ 112

Procedures ..................................................................................................................... 113

Findings ......................................................................................................................... 114

Discussion ..................................................................................................................... 120

The Culture of Research and Evaluation ..................................................................... 123

Implications for Informed Consent Practices ............................................................. 124

Conclusion ..................................................................................................................... 126

V. CONCLUSION .......................................................................................................... 129

Overview ...................................................................................................................... 129

Paper 1 Summary ........................................................................................................ 130

Conclusions ................................................................................................................. 130

Limitations .................................................................................................................... 130
Table of Contents - continued

CHAPTER

Future Research ................................................................. 133

Paper 2 Summary ................................................................. 133

Conclusions ........................................................................ 133

Limitations .......................................................................... 134

Future Research ................................................................. 135

Paper 3 Summary ................................................................. 136

Conclusions ........................................................................ 136

Limitations .......................................................................... 137

Future Research ................................................................. 138

Contribution to Evaluation .................................................. 139

Contribution to Theory ......................................................... 139

Contribution to Methodology and Practice ......................... 140

REFERENCES ...................................................................... 142

APPENDICES ....................................................................... 159

A: Paper 1 Survey Sample .................................................. 159

B: Paper 2 Survey Sample .................................................. 162

C. HSIRB Approval Scans (Papers 1-3) .............................. 166
LIST OF TABLES

1. Focal Research Questions and Anticipated Contributions to Evaluation........ 12
2. Literature Review Themes: Visual Methodology ......................................... 14
4. Overview of Paper Methodology .................................................................. 21
5. Participant Demographics ........................................................................... 42
6. User-Generated Product Review Videos Coded by Study Participants .......... 44
7. Word Count Comparison ............................................................................. 46
8. Mturk and Evaluator Qualitative Comparison ............................................. 48
9. User-Generated Product Review Videos Coded by Study Participants .......... 75
10. Participant Demographics as Percentage of Cohort ................................... 76
11. Overview of Average Agreement by Cohort .............................................. 82
12. 3 Coder Cohort Agreement ......................................................................... 83
13. 10 Coder Cohort Agreement ....................................................................... 84
14. 20 Coder Cohort Agreement ....................................................................... 85
15. Average Agreement by Video and Coder Cohort ....................................... 86
16. Correlation Between Personal Relevance and Emotive Coding ................. 87
17. Data Categories and Description .................................................................. 112
I. INTRODUCTION

Photography was invented in 1798 (Gernsheim, 1986); Thomas Edison patented the first movie camera in 1888 (Green, 2013); and methods for capturing still and moving images have been progressively refined and simplified ever since. This trend especially accelerated in recent years with low-cost digital cameras on cellphones; by 2010, over one billion camera phones were in use worldwide (The Economist, 2010). As just one indicator for the total volume of photos taken, 300 million photos are uploaded to Facebook every day, and the total archive on Facebook currently numbers 240 billion photos (Madrigal, 2013). In 2008 users were viewing photos at a rate of 300,000 per second, so conservative estimates are that several million photos are viewed each second on Facebook (Madrigal, 2013). In addition to those substantial numbers, the fact that Facebook was launched as a student project in just 2004 (Tabak, 2004) underscores the speed of change and broad demand for these forms of communication. As for video, YouTube was launched in 2007 and currently has over 1 billion unique visitors a month, 70% of which are outside the United States. These visitors watch 6 billion hours of video each month, and upload over 100 hours of new video content every minute (YouTube, 2013). These historical facts and statistics underscore two important points related to the contextual framework of this dissertation:

1) There is over 120 years of history of recording still and moving images. Even though relatively recent technologies have exponentially increased
the accessibility of these visual communications. These developments were expansions or modifications of long-established methods, thus are not ‘new.’ Similarly, the core issues of utilizing visual methodology in research and evaluation such as ethics, data collection, or analysis are not ‘new’ (Shrum, Duque & Brown, 2005, Gibbs et al., 2002, Banks, 2001).

2) Visual communications (photographs and video) are an established part of the cultural context for a large number of people, especially in the United States and other developed countries. Given this fact, “it would be ethically dubious, perhaps even academically negligent, nowadays to overlook the visual enskilment of the greater scholarly and general population” (Perry & Marion, 2010, p. 99). At a minimum, evaluators should be aware of these trends and consider the implications for their work in order to uphold the principle of cultural competence for evaluators (American Evaluation Association, 2009).

**Context of Complexity**

The problem addressed by this study was developed within both the historic and cultural context related to the visual data noted above as well as the conceptual context of evaluation under conditions of complexity. The broad need for new evaluation methodology and theory within this domain was succinctly stated by Patton: “Evaluation has explored merit and worth, processes and outcomes, formative and summative evaluation; we have a good sense of the lay of the land. The great
unexplored frontier is evaluation under conditions of complexity” (Patton, 2011, p. 1). A primary reason complexity remains an ‘unexplored frontier’ is the conceptual legacy of a reductionist scientific paradigm which engendered the belief that “we could study the parts…and arrive at knowledge of the whole. We have reduced and described and separated things into cause and effect, and drawn the world in lines and boxes” (Wheatley, 1999, p. 29). While this paradigm has served the evaluation community well by providing such useful tools as random control trials and logic models (Mathison, 2005), other tools are needed to address the problems within a context of complexity (Patton, 2011).

For the purposes of this dissertation, ‘complexity’ is understood as “characterized by a large number of interacting and interdependent elements in which there is no central control” (Mitchell, 2009, p. 13). Instead of centralized control, these ‘interacting and interdependent elements’ operate in a nonlinear system which is ‘webbed with feedback loops’ which feed change back on itself and amplify or otherwise inform ongoing growth and change (Wheatley, 1999). This context is particularly challenging for evaluators (i.e., Pawson, Wong, & Owen, 2011), because “in a nonlinear world, very slight variances, things so small as to be indiscernible, can amplify into completely unexpected results...after several iterations, a variance that was too small to notice can cause enormous impact, far beyond anything predicted…in a nonlinear world, there is no relation between the strength of the cause and the consequence of the effect” (Wheatley, 1999, p. 136-137, emphasis added). An additional difficulty in approaching this problem is the fact that evaluation is itself part
of the complexity of human decision making and information feedback, “just one part of a complex, interdependent, nonlinear set of problem-solving activities” (Shadish, Cook, & Leviton, 1991, p. 21).

Just as ‘fish find water last,’ human life is lived in an ocean of complex information and feedback systems, much of which is essentially evaluative. Most advertising (another significant component of the visual cultural context in much of the world) centers on evaluative claims such as the ‘best’ furniture value, ‘most nutritious’ breakfast cereal, ‘most effective’ detergent, etcetera. Similarly, many conversations with friends and family revolve around evaluative themes, such as the ‘most enjoyable’ vacation, ‘best’ restaurants, or ‘easiest’ way to potty-train a child. Beyond these domestic examples, complexity pervades most programs related to human change, growth or development – which is to say, a high percentage of the total. “Social programs are undeniably, unequivocally, unexceptionally social systems, and they are composed, as is any social system, of the interplay of individual and institution, of agency and structure, of micro and macro social processes. Much is to be learned from inspecting the ‘social nature’ of programs” (Pawson, & Tilley, 1997, p 406). Two well-known examples include Weight Watchers and Alcoholics Anonymous, which have built their programs around delivering personalized evaluative feedback that is especially potent because it is delivered in the context of personal relationships developed in regular small group meetings.

Since complexity pervades both the routine and programmed aspects of human lives, evaluators must develop methods that work within this context…especially if
they are to “help to answer big questions of public management and public administration” (Wholey, 1997, p. 132). Advances toward this end have been made by systems thinking (i.e., Hargreaves & Podems, 2012; Williams, 2008) as well as complex causality (i.e., Sager, & Andereggen, 2012), and within the evaluation community, especially by the book Developmental Evaluation (Patton, 2011). Both the articulation of the challenges associated with complexity and the initial theories and insights these sources have provided comprise an important aspect of the context for this dissertation.

**Benefits of Visual Methods**

Many of the benefits of visual methods flow from their unique emotional and cognitive impact. This combination of ‘head’ and ‘heart’ make them an especially significant and emotionally relevant form of communication (Oware, Diefes-Dux & Adams, 2007).

“The visual has an explicitness and immediacy which delivers a multisensory impact. This immediacy of the visual affects us in a profound and elusive way – before the sense-making apparatus, the cognitive processing, there is a pre-reflective reaction, as several writers and researchers have noted. There seems to be some accord that there is something indefinable about the visual, grounding it in material reality. It is an immediate and authentic form which verbal accounts are unable to fully encompass” (Spencer, 2011, p. 32).
This unique cognitive impact is such that images have a distinctive function in bridging psychological and physical realities such as emotional reactions or memories associated with images of specific locations or people. Because of this unique strength, “photographs appear to act as both stimuli and verifiers of perception” (Tucker & Dempsey, 1991, p. 649). Evaluators have used an interview method that uses photographs instead of text prompts (also known as ‘photolanguage’ or ‘photelicitation’) to harness this strength (i.e., Bessell, Deese, & Medina, 2007; Oware, Diefes-Dux, & Adams, 2007). To illustrate the unique power of photos as a prompt for discussion, another researcher noted how “photographs also led to some fascinating discussions about wider issues relating to politics, identity, racialised politics, aspirations, music, sexuality, and role-models…there was no way I could conjure up a set of questions to elicit from that young person such a profound discussion” (Clark, 2010, p. 405).

Given the complex feedback of human communication (noted in the context section above), “images and video open up complex, reflexive and multi-faceted ways of exploring social realities” and provide a uniquely powerful form of ‘thick description’ (Spencer, 2011, p. 34). Visual data is considered especially ‘rich’ when used in providing ‘thick description’ (Oware, Diefes-Dux, & Adams, 2007) because it contains and communicates a uniquely large amount of information in an efficient format – hence the cliché that ‘a picture is worth a thousand words.’ Because of the unique strengths of visual communications, visual records are especially useful for research and evaluation. This point was highlighted by a social science researcher that
noted “watching ‘real’ people from different communities talk about their lived experiences...captures an embodied expression, not abstract truisms…” (Spencer, 2011, p. 33). Similarly, the potential impact of photographs in communicating findings (Jacobs, 1999) and especially shaping ethical behavior is powerfully illustrated by the documentary photo taken during the Vietnam war of the naked girl running down the street after a napalm attack, which “may have done more to halt the Vietnam War than all the writings of moral philosophers of the time put together” (Blackburn, 2002, p. 5).

Visual data also provides unique benefits related to the typically deeper involvement of participants in the process of collecting the data (Shrum, Duque, & Brown, 2005) and in the collaboration with other researchers or colleagues during data analysis phases (Luff & Heath, 2012). This strengthened partnership and collaboration improves the rigor of the study by increasing transparency and bringing additional insights of other perspectives:

“The value of showing and sharing data with colleagues and peers should not be underestimated. A long-standing criticism of ethnography concerns the lack of its ‘transparency’; critics highlight the difficulties of recovering what the researcher saw and experienced undermining the ability of fellow scholars to form an independent judgment of the quality of the analysis” (Heath, Hindmarsh, & Luff, 2010, p. 7).
Benefits of Video Data

In addition to the benefits from visual methods in general, unique benefits related to the use of video data are also prominent in the literature, and are therefore especially relevant to this dissertation. In evaluation, while relatively rare, video data was cited as being used to help estimate the number of homeless people in an urban context (Berry, 2007), and in evaluations of educational programs (Ingle, 1984; Hurworth & Sweeney, 1995). Additionally, it was encouraged for consideration in any study using a mixed methods design (Bennington, Gay, & Jones, 1999). In qualitative research more broadly, video-based studies were referenced in a range of areas, including medicine (Guerlain, Calland, Adams, & Turrentine, 2004), work, and organization (Fele, 2012), interpersonal communication, learning, human interaction with technology, the family and the home, and professional practice – especially related to training and communications (Heath, Hindmarsh, & Luff, 2010).

One of the primary benefits of video data during the analysis phase is that it provides the opportunity for repeated viewings of key activities of the evaluand (such as classroom behavior or teaching techniques) and can facilitate a deeper understanding and analysis (Jacobs, 1999; Haidet, Tate, Divirgilio-Thomas, Kolanowski, & Happ, 2009). In this sense, video extends the afterlife of research or evaluation experiences as it provides a uniquely rich source of data to help the researcher or evaluator relive the experience under scrutiny (Shrum, Duque, & Brown, 2005; Mondada, 2006). Additionally, “video recordings are an excellent source of data
that can be used to assess relationships between behaviors that occur in close temporal proximity to one another, and provides a high degree of reproducibility when measuring observations” (Haidet et al., 2009).

From the perspective of a constructivist epistemology, video is especially beneficial as it “offers a ‘microscope’ for an in-depth study of the on-going production of situated social order” (Knoblauch & Schnettler, 2012, p. 335) and helps “to take seriously the ways in which social action is produced as intelligible by virtue of the interplay of the spoken, the visible and the material” (Heath, Hindmarsh, & Luff, 2010, p. 146).

Finally, the benefit most relevant to the selection of video data for this dissertation is that it provides a uniquely powerful feedback mechanism and communication vehicle to ‘harness complexity’ and inform human development (Patton, 2002) and complex organizations (Heath, Hindmarsh, & Luff, 2010, p. 143). Video data provides a powerful feedback and training tool (Mackenie, 2004, p. 195) that provides the ‘double-loop’ learning needed to prevent the problem or embed the solution in a changed system (Patton, 2011, p. 11). For example, in a medical context it was found that “using patients’ images and voices to communicate quality improvement opportunities, can provide greater depth of understanding than traditional interviews, focus groups, or surveys…The voices of real patients in videos motivate change in ways that other data cannot. Through this approach, we have been able to uncover the hidden reasons behind events, explain discrepancies between what
people say and what they do, and identify needs that our patients can’t always articulate” (Kaiser Permanente, 2010, p. 2, emphasis added).

Problem

The four-part context described above (historic, cultural, complexity, and value of visual methods) touch themes that are very broad. Within that context, the problem identified in the literature review below is specific: the underutilization of visual methods within the field of evaluation. The goal of this dissertation is to contribute knowledge and insight into evaluative work with visual data and practical resources for evaluators (and potentially social science researchers) to address this underutilization. This underlying goal was pursued through three papers (further described in the paper summary section at the end of this chapter):

1. A comparative analysis of the open-coding provided through Amazon’s Mechanical Turk (Mturk) crowd-sourcing platform with a control group of evaluators.

2. An analysis of the reliability of categorization coding by three sizes of groups of Mturk participants; 3, 10 and 20 participants.

3. A conceptual framework for addressing unique ethical challenges posed by visual methods, as fear over ethical concerns and restrictions or inappropriate parameters of Institutional Review Boards are cited as a barrier to use of visual
methods (i.e., Allen, 2009; Wiles et. al, 2008; Emmison & Smith, 2000; Perry & Marion, 2010).

Two design features of papers 1 and 2 are important relative to the problem statement, because these features are the basis for the relationship between the specific problem and broad themes and goals noted above. First, the data used for papers 1 and 2 is user-generated videos hosted on the YouTube channel (Berthon, DesAutels, & Pitt, 2011) of ExpoTV. While user-generated evaluation videos are relatively uncommon, these videos have been demonstrated as providing clear evaluative feedback and conclusions (Wilson, 2011) and can therefore be treated as a source of evaluative data. As such, these product review videos are part of the broader category of ‘crowd-sourced evaluations’ that Michael Scriven (2013) noted should be “treated as a valuable - indeed often invaluable - source of information…” (p. 3-4). Second, out of the broad range of data utilized within the domain of visual methods, video data was specifically selected because within a complex system involving multiple feedback mechanisms, video communications are uniquely powerful for communication and training (Mackenie, 2004). While other evaluators have undoubtedly benefited from the additional value video data provides, Michael Patton (2002) specifically referenced using video data to provide visual feedback to staff (p. 308) and noted how this form of data could be useful in a variety of contexts as “sometimes videotapes originally done for research or evaluation can subsequently be used for future training, program development, and public relations, making the costs more manageable because of added uses and benefits” (p. 308). A summary of the
focal research questions and anticipated contributions to the field of evaluation is provided in Table 1 below:

Table 1
Focal Research Questions and Anticipated Contributions to Evaluation

<table>
<thead>
<tr>
<th>Category</th>
<th>Paper 1</th>
<th>Paper 2</th>
<th>Paper 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal Research Question 1</td>
<td>How does crowd-sourced open coding of video data compare with coding provided by trained evaluators?</td>
<td>What is the cost, speed, and accuracy of crowd-sourced categorization coding of video data?</td>
<td>What are the assumptions and procedures guiding the use of images guiding the use of images within the fields of entertainment, journalism, advertising and the internet?</td>
</tr>
<tr>
<td>Focal Research Question 2</td>
<td>None: Paper 1 had only one focal research question.</td>
<td>How do different types of categorization questions and different sizes of coder cohorts impact the results?</td>
<td>Can broad principles or themes from these procedures be identified to provide a contextually appropriate reference for guiding ethical decisions within the framework of ‘relational integrity’?</td>
</tr>
<tr>
<td>Anticipated Contribution to Evaluation #1</td>
<td>Demonstrate how crowd-sourced individuals are uniquely positioned to analyze user-generated evaluation data; the crowd should interpret the crowd.</td>
<td>Pilot test a new analytic method for categorical coding of video data that specifically address key barriers to visual methodologies of time and cost.</td>
<td>Expand and clarify the theme of valuing relationship with participants found in literature; ‘relational integrity’ is foundation for ethical visual methodologies.</td>
</tr>
</tbody>
</table>
Table 1 - continued

<table>
<thead>
<tr>
<th>Category</th>
<th>Paper 1</th>
<th>Paper 2</th>
<th>Paper 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anticipated Contribution to Evaluation #2</td>
<td>Pilot test a new analytic method for open coding of video data that specifically address key barriers to visual methodologies of time and cost.</td>
<td>Demonstrate high reliability of crowdsourced coding under specific circumstances; outline additional research to further develop method.</td>
<td>Provide resource for evaluators and IRB committees by outlining four categories of relationship from other industries.</td>
</tr>
<tr>
<td>Anticipated Contribution to Evaluation #3</td>
<td>Demonstrate a method to realistically gather exponentially more perspectives on qualitative data than standard; superior to post-analysis 'member checks' as representatives of member groups directly provide analytic perspective.</td>
<td>Outline specific applications of method to serve as resource for practicing evaluators and inspiration for additional research.</td>
<td>Provide resources and context to address ethical fears as contribution to overall goal of reducing barriers to the use of visual methodologies (i.e., time/cost of analysis from papers 1 and 2).</td>
</tr>
</tbody>
</table>

Literature Review

Overview

A literature review was conducted to provide grounding and context for this study, and included both a search of relevant key words in the primary evaluation journals (further described in the ‘visual methods in evaluation’ section below) and a broad search of social science literature. This search included broad terms related to
visual methodology and visual ethics on Google Scholar and the ‘PowerSearch’ of the WMU Library, as well as specific searches for resources referenced in the bibliographies of seminal articles and books. In order to systematize the process, notes from book sources were typed in Word and an Excel spreadsheet was developed with relevant excerpts from the journals and internet resources. These excerpts were coded by theme, and Table 1 below provides a high-level overview of the findings of the literature search, listing both the primary themes found in the literature as well as the number of excerpts coded per theme as an indicator of prevalence in the literature.

Table 2

<table>
<thead>
<tr>
<th>Category</th>
<th>Theme</th>
<th>Excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>Video methods</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Photo methods</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>New Methods</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Qualitative methods</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Qualitative research</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Visual methods</td>
<td>20</td>
</tr>
<tr>
<td>Ethics</td>
<td>Visual Ethics</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Photo method ethics</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Qualitative method ethics</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Contextual Integrity</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Visual Ethics History</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Informed consent</td>
<td>10</td>
</tr>
<tr>
<td>Mturk</td>
<td>Mturk overview</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Mturk data</td>
<td>10</td>
</tr>
<tr>
<td>Analysis</td>
<td>Video analysis</td>
<td>42</td>
</tr>
</tbody>
</table>
Table 2 – continued

<table>
<thead>
<tr>
<th>Category</th>
<th>Theme</th>
<th>Excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative Visual</td>
<td>Qualitative Visual analysis</td>
<td>42</td>
</tr>
<tr>
<td>Qualitative Rigor</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Complexity</td>
<td>System dynamics</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Evaluating Communication</td>
<td>6</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>743</td>
</tr>
</tbody>
</table>

*Visual Methods in the Social Sciences*

The use of visual methods within the social sciences more broadly defined was considered an appropriate element of the context for this evaluation dissertation because many evaluation methods emerged from social science research traditions (Worthen, Sanders, & Fitzpatrick, 1997, p. 4). For the purposes of this study, the term ‘visual methods’ refers to the qualitative research methodology involving the collection and analysis of visual data, especially focused on the categories of ‘researcher created data’ (i.e., researcher-generated video or photography) and ‘respondent created data’ (i.e., user-generated video or photography), though it is recognized that ‘found data’ (i.e. postcards, web images) and ‘representations’ (i.e., models or graphics from a variety of creative sources) are also included in this broad classification (Prosser & Loxley, 2008, Wiles et al., 2008).

While some might consider visual methods to be ‘new’ approaches to research, visual methods actually have a long history within the social sciences. For example, visual ethnography developed alongside documentary photography and film-making which began in the Victorian era of the late nineteenth century (Shrum, Duque &
Brown, 2005), and photographs and film have been used more widely in sociology and anthropology since the 1920’s (Gibbs et al., 2002). As an established specialty in social research (Banks, 2001), the community surrounding the use of visual methods has organized both the International Visual Sociology Association (IVSA) and Commission on Visual Anthropology (CVA). The IVSA hosts annual conferences and publishes a peer-reviewed journal called *Visual Studies* (IVSA, 2013), and the CVA hosts its own program within the conferences of the International Union of Anthropological and Ethnological Sciences and publishes *Visual Anthropology* (CVA, 2010). In addition to these branches within ethnography, sociology and anthropology, visual methods have been used in the disciplines of communication, education, photojournalism, cultural studies, ethnic studies and industrial management (Harper, 2005). Furthermore, they have been an ongoing aspect of studying the human factors related to workplace training and observation, a usage that has especially accelerated with the advent of digital imaging (Guerlain, Calland, Adams, & Turrentine, 2004).

Additionally, visual methods have been used in scientific research related to resource management and agricultural applications since at least the 1980’s, as well as in studies related to damage assessments after natural disasters (Kerle, Stekelenburg, Van den Heuvel, & Gorte, 2005).

*Underutilization of Visual Methods in Evaluation*

Within the evaluation community, the use of visual methods has “had a chequered history and so interest seems to have moved through peaks and troughs”
with the most recent spike in interest being in the 1980’s (Hurworth, & Sweeney, 1995, p. 153). For example, in a 1985 article in the American Journal of Evaluation, Fang suggested a regular use of photos, saying “photographs serve as another source of data, just as interviews, observations, and written evaluation forms do” (Fang, 1985, p. 24).

In order to assess the scope and context of the use of visual methods within evaluation methods and theory, the literature review for this dissertation included a search of the four primary evaluation journals, including the American Journal of Evaluation (AJE), New Directions in Evaluation (NDE), Evaluation Review (ER) and the Journal of Multi-Disciplinary Evaluation (JMDE). The search included 11 words or phrases ranging from broad (i.e., ‘photo’ or ‘video’) to specific (i.e. ‘video data’ or ‘visual survey’) in addition to two phrases related to the crowd-sourcing methodology used in papers one and two of this study, and the results are presented in Table 2 below:

Table 3
‘Visual Methods’ Literature Search of Evaluation Journals

<table>
<thead>
<tr>
<th>Keyword</th>
<th>AJE</th>
<th>NDE</th>
<th>JMDE</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>63</td>
<td>28</td>
<td>7</td>
<td>42</td>
</tr>
<tr>
<td>&quot;Video Analysis&quot;</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&quot;Video data&quot;</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Photo</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>&quot;Photo-elicitation&quot;</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&quot;Visual data&quot;</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>&quot;Visual methods&quot;</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Visual survey&quot;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&quot;Image use&quot;</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
The first general finding from this search, consistent with the reference noted above, is that visual methods are an established part of evaluation practice, as the search identified articles referencing visual methodologies from the 1980’s to the present (content description and citations below). However, deeper analysis of the literature revealed a second and equally clear finding: visual methods are a relatively rare methodology within the field of evaluation. For example, of the 140 articles including the term ‘video’ in the journals noted above, none published in the AJE, JMDE or ER used the word in the title, and only 1 article published in the NDE included the word in the title: “Microcomputers in schools: The video case study as an evaluation tool” (Ingle, 1984). Similarly, when searching the abstracts of the articles there was no mention of the term ‘video’ in AJE, only 2 in NDE (Ingle, 1984, Bennington, Gay, & Jones, 1999), 1 from JMDE (Wilson, 2011) and 1 incidental reference in an abstract in ER. Likewise, there was no reference to the word ‘photo’ in any of the NDE articles and neither the title nor abstract in any of the AJE or JMDE articles. The one reference in a title of an ER article was “Photo-Interviewing: A Tool for Evaluating Technological Innovations” (Tucker, & Dempsey, 1991). The remaining references to these keywords were not related to substantive use of visual

Table 3 – continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>AJE</th>
<th>NDE</th>
<th>JMDE</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowdsource</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Crowd-sourced</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
methods but rather incidental references such as to a video cassette, a photo, a video curriculum used by the evaluand, or YouTube videos.

Despite the “rhetoric espousing the idea that data in the form of the visual image would be beneficial for evaluations generally” (Hurworth, & Sweeney, 1995, p. 153), there was also “little recent evidence in evaluation literature to suggest either the value of the visual image within data management or that photographs and videos can be used in a variety of evaluation designs” (Hurworth, & Sweeney, 1995, p. 153). This same article concluded with the enthusiastic prediction that “in the very near future the technology to scan images into computer programs and then to sort material in multiple ways will become more commonplace and will make it more plausible and exciting to include visual images in the evaluation process” (Hurworth, & Sweeney, 1995, p. 163). While there have indeed been broad advances in technology over the 12 years since this prediction was made, the finding of the evaluation literature search is that the use of visual methods in evaluation is still remarkably rare.

*Underutilization of Visual Methods in the Social Sciences*

The literature review also found several references to underutilization within the broader social sciences as well, and the reasons for this broader underutilization helped frame the studies for this evaluation-specific dissertation. One article noted that “visual methods are marginalized as forms of data collection within the field of education and when used, have been employed disparately” (Allen, 2009, p. 397). Another article referenced video specifically as “under-utilised in the social sciences
as a means of gathering data concerning everyday social interaction. This is despite innovations over a century ago in adopting video, or more accurately the moving image, to understand naturalistic conduct” (Luff & Heath, 2012, p. 257). The book Video in Qualitative Research (Heath, Hindmarsh, & Luff, 2010) opens by several references to the underutilization of video, including:

- “while audio-visual recordings provide unique access to the details of social action, they are relatively under-utilised in the social sciences despite their significant potential…” (p. 1-2)
- “research project after research project fail to include filming and insist on continuing the hopelessly inadequate note-taking of an earlier age” (1995 quote form Margaret Mead, p. 2)
- “The seeming neglect of video and the moving image in the social sciences is particularly curious when we consider that soon after the development of instantaneous photography in the 1830’s, implications for the analysis of human and behavioural sciences was recognised” (p. 3).

A primary conceptual reason for this underutilization is that what has typically been understood as ‘social science’ has been framed as involving verbal or written language (Allen, 2009, p. 397), such that visual methods have been considered an “unconventional, ‘isolated’ and ‘somewhat eccentric specialism…a symptom of these conceptual shortcomings is the widespread tendency to use visual materials in a purely illustrative, archival or documentary way rather than giving them a more analytic treatment” (Emmison & Smith, 2000, p. ix). In addition, there are several practical
reasons for the underutilization. First, these methods are subject to the sometimes steep learning curve associated with any technology (Travers, 2009). Additionally, visual methods are a time-consuming method of data collection (Fang & Ellwein, 1990), and camcorders in particular are perceived as “a more intrusive technology, a more threatening character, a more engaging actor on the stage” (Shrum, Duque, & Brown, 2005, p. 9). Finally, while visual methods offer various strengths, (described below), another reason for underutilization is that they do little to change the fundamental practical considerations faced by social science researchers such as budget, schedule and access to participants, nor do they alter the core tensions surrounding theoretical choices (Travers, 2009, Luff & Heath, 2012).

Table 4

*Overview of Paper Methodology*

<table>
<thead>
<tr>
<th>Category</th>
<th>Paper 1</th>
<th>Paper 2</th>
<th>Paper 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>143 total surveys with descriptive open codes of video data</td>
<td>169 total surveys with categorization codes of video data</td>
<td>Image use forms and procedures intentionally sampled from fields of entertainment, journalism, advertising and social media</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>5 internet surveys distributed through Mturk; sample included as Appendix A</td>
<td>5 internet surveys distributed through Mturk; sample included as Appendix B</td>
<td>None</td>
</tr>
</tbody>
</table>
Table 4 – continued

<table>
<thead>
<tr>
<th>Category</th>
<th>Paper 1</th>
<th>Paper 2</th>
<th>Paper 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>*Treatment: 106 unique Mturk participants</td>
<td>103 unique Mturk participants</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>*Control: “Typical” qualitative analysis team; 3 trained evaluators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>Qualitative content analysis</td>
<td>Fidelity analysis (percentage), phi coefficient</td>
<td>Two stage coding and inductive analysis of underlying assumptions</td>
</tr>
</tbody>
</table>

Paper 1 Summary

*Title: Harnessing Complexity: Crowd-Sourced Open-Coding of User-Generated Evaluation Data*

This study addresses the underutilization of visual data in evaluation by using Amazon’s Mechanical Turk (henceforth ‘Mturk) crowd-sourcing platform for analyzing video data. The crowd-sourcing aspects of this study were built on the foundation of work by Azzam and Jacobson (2013) related to the potential value of utilizing Mturk in various elements of evaluation projects. Additionally, this study addresses the question Dr. Azzam posted on the AEA 365 blog about exploring the use of Mturk for qualitative analysis (Azzam, 2013). Furthermore, this study builds on
others which utilized crowd-sourced workers for analysis (i.e. Bohannon, 2011; Little, Chilton, Goldman, & Miller, 2009; Downs, Holbrook, Sheng, & Cranor, 2010).

The focal research question for paper 1 is: How does crowd-sourced open coding of video data compare with coding provided by trained evaluators? This study is specifically addressed toward crowd-sourcing the “open coding” (Corbin & Strauss, 2008) step of the analysis, part of the initial inductive phase of analysis where “findings emerge out of the data…in contrast to deductive analysis where the data are analyzed according to an existing framework” (Patton, 2002, p. 453). The data which is analyzed by the two groups is user-generated video data from ExpoTV. In addition to the benefit of working with crowd-sourced evaluation data (as noted above), this also demonstrates the analytic potential for material typically relegated only to illustration or documentary purposes (Emmison & Smith, 2000, p. ix).

Five internet surveys (one for each product review video) were developed in SurveyMonkey to efficiently facilitate the involvement of the geographically diverse participants serving in the two groups compared in this study. The video data was directly embedded in these surveys, which also included prompts for the open coding and descriptive paragraphs and demographic data, and a sample is included in Appendix A. Links to these surveys were distributed to the Mturk participants utilizing the Mturk platform and to the evaluators by emailed links. The qualitative analysis conducted for this study is within the broad category of “content analysis” (Patton, 2002, p. 453), such that the prompts on the surveys were to “describe the content” in both 30 increments and as an overall descriptive paragraph. Consistent
with the content analysis approach, the analysis of the analytic data provided by study participants includes both word counts and analysis of broad categories or themes (Marshall & Rossman, 2011, p. 161).

The findings indicated that Mturk participants provide more analytic data than required by the internet survey instrument used to facilitate their participation, and provided both detailed descriptive codes and candid overall impressions of the video data. A primary quantitative finding in the comparison between the Mturk participants and evaluator control group was that the average words provided per second of video data decreased for the evaluators over the course of the five videos they reviewed, while the Mturk participants provided consistent quantity throughout. A primary qualitative finding in the comparison was that the evaluator coding and descriptions were broad, professional and evaluative, while the Mturk participants were much more detailed, descriptive, and unprofessional to the point of crass in some cases. The broad conclusions from this study indicate that while there are specific limitations, the Mturk methodology offers considerable potential advantages for evaluators: the broader perspective available through a diverse group of participants provides a uniquely useful source for relevant analytic data that is both time efficient and cost effective.

**Paper 2 Summary**

*Title: Crowd-Sourced Video Data Coding: Exploring new opportunities and tradeoffs utilizing the Mturk platform*
This exploratory study was developed as a continuation of the study outlined in Paper 1 above, and both were built on the foundational work of Azzam and Jacobson (2013) to explore the use of crowd-sourcing in evaluations. Both studies utilize the same user-generated product review videos as the underlying data for crowd-sourced analysis, such that the findings of the open-coding study served as a foundation for development of the current study by guiding the formulation and description of the categories for coding.

The primary questions guiding this exploratory study are: what is the cost, speed and accuracy of crowd-sourced categorization coding of video data, and how do different types of categorization questions and different sizes of coder cohorts impact the results? Similar to paper 1, this study provides as context an overview of the history, strengths and limitations of the Mturk crowd-sourcing platform, and a range of studies utilizing crowd-sourcing for research and evaluation are cited. In order to facilitate a comparative analysis, the categorical coding was completed by cohorts of 3, 10 and 20 coders, and these participants represented considerable demographic and geographic diversity.

The findings indicated that the coding provided by crowd-sourced participants is nearly perfectly accurate for the presenters gender, is generally high for objective details present in the video content, and shows mixed degrees of accuracy for a range of other coding decisions. While there are specific limitations and further study is required to further refine the methodology, the broad conclusions from this study indicate that this methodology does provide a potential to reduce barriers to visual
methodologies. For example, coding for video data can be completed in a matter of hours at a reasonable cost, and specific opportunities for applying this method include initial data description and sorting, analyzing existing video data, and facilitating the analysis of new user-generated video evaluation data.

Paper 3 Summary

Title: Relational Integrity: A Conceptual Framework for Ethical Visual Methodology Derived from the Fields of Entertainment, Journalism, Advertising and Social Media

Using photographs or video as data raises ethical questions many evaluators and researchers feel unprepared to address, and this hesitation and confusion is a barrier to the more widespread use of this important source of data (Allen, 2009, Perry & Marion, 2010). The theme of ‘relational integrity’ is identified in the literature of ethical visual methodologies, and is suggested as a conceptual frame that compliments the Code of Research Ethics and Guidelines published by the International Visual Sociology Association (Papademas, 2009). The implications and potential benefits of this approach are explicated for the especially challenging ethical domains of informed consent and confidentiality/anonymity, with particular focus on the parallels between ‘relational integrity’ and ‘contextual integrity’ which has been advanced within the legal community (Nissenbaum, 2004).

Key themes and findings from a comparative analysis of the image use procedures in the fields of entertainment, journalism, advertising, and social media are presented as contextual reference for guiding ethical decisions and informing
relationships developed surrounding visual methodologies. Five categories of relationship are suggested based on this study, including recruited individual, employed individual, event attendee, anonymous public, and community participant. Suggested applications of the ‘relational integrity’ conceptual framework and findings from the comparative analysis in varied evaluation and research projects are provided, including guidance on how to develop and maintain consent forms within an overall relationship-building context, ensuring that any potentially objectionable research needs are communicated in writing in advance, and keeping informed consent in the parameters appropriate with the level of relationship.
II. CROWD-SOURCED OPEN CODING OF USER-GENERATED EVALUATION DATA

Authors

Kurt Wilson, Western Michigan University and Tarek Azzam, Claremont Graduate University

Abstract

This study explores the potential uses of crowd-sourced data analysis methods for qualitative, open-coding analysis of video data in evaluation. An overview of the history, strengths and limitations of crowdsourcing research and evaluation projects is provided, and a range of studies utilizing similar methodologies are cited. The video data coded was user-generated video reviews of various products, and coding from crowd-sourced participants was compared with coding from a control group of trained evaluators.

The findings indicated that crowd-sourced participants provide more analytic data than required, and provided both detailed descriptive codes and candid overall impressions of the video data. The broad conclusions from this study indicate that while there are specific limitations, the crowdsourcing methodology offers considerable potential advantages for evaluators, including saving time and money in coding video data and benefiting from the analytic perspective of a broader group of participants.
Keywords: Video data, qualitative analysis, Mturk, crowd-sourced, user-generated, product review video, ExpoTV, evaluation

Introduction

The primary tools of visual communication have a long history: photography was invented in 1798 (Gernsheim, 1986) and Thomas Edison patented the first movie camera in 1888 (Green, 2013). More recent advances in digital imaging and social media have built on this established foundation, accelerating the trends of lowering costs and broadening usage. For example, over one billion camera phones were being used across the world by 2010 (The Economist, 2010). Facebook statistics provide useful indicators for the vast numbers of digital photographs both produced and consumed: currently 300 million photos are uploaded to Facebook every day, and the total archive on Facebook currently numbers 240 billion photos (Madrigal, 2013). As related to photo viewing, in 2008 users were viewing photos at a rate of 300,000 per second, so conservative estimates are that several million photos are viewed each second on Facebook (Madrigal, 2013). In addition to those substantial numbers, the fact that Facebook was launched as a student project in just 2004 (Tabak, 2004) underscores the speed of change and broad demand for these communications. As for video, YouTube was launched in 2007 and currently has over 1 billion unique visitors a month, 70% of which are outside the United States. These visitors watch 6 billion hours of video each month, and upload over 100 hours of new video content every minute (YouTube, 2013). Finally, these ‘user-generated’ visual communications are
just a small part of the total visual environment most people are immersed in, as the professionally produced advertising, magazines, newspapers, feature films and television comprise a major element of the overall total.

Clearly, visual communications are a substantial element of global culture and important element in human communications, especially in developed countries. Given both the richness of visual data and principle of cultural competence for evaluators (American Evaluation Association, 2009) the underutilization of visual methods in the evaluation community is unfortunate; one source went so far as to say, “it would be ethically dubious, perhaps even academically negligent, nowadays to overlook the visual enskilment of the greater scholarly and general population” (Perry & Marion, 2010, p. 99). The exceptions within the past thirty years of evaluation literature (e.g., Ingle, 1984; Fang, 1985; Bennington, Gay, & Jones, 1999; Tucker, & Dempsey, 1991; Hurworth, & Sweeney, 1995; Bessell, Deese & Medina, 2007; Oware, Diefes-Dux, & Adams, 2007) are notable in their rarity. In the broader research community, visual methods are also relatively marginalized, but nonetheless are an established specialty (Banks, 2001) represented by both the International Visual Sociology Association (IVSA) and Commission on Visual Anthropology (CVA), each of which host their own conferences and journals, Visual Studies (IVSA, 2013), Visual Anthropology (CVA, 2010).
**User-Generated Video Data**

User-generated videos such as those posted on YouTube are examples of the increasing cultural significance of visual communications. While there are few published examples of this type of data being used in program evaluations, this category includes the many product review videos or patient testimonials posted on YouTube. Evaluative videos of this nature are part of the growing body of crowd-sourced evaluation data which includes Amazon or Zagat reviews, which should be “treated as a valuable—indeed often invaluable—source of information” (Scriven, 2013, p. 3-4).

Video is an especially powerful form of communication within complex systems because the “immediacy of the visual affects us in a profound and elusive way – before the sense-making apparatus, the cognitive processing, there is a pre-reflective reaction, as several writers and researchers have noted. There seems to be some accord that there is something indefinable about the visual, grounding it in material reality. It is an immediate and authentic form which verbal accounts are unable to fully encompass” (Spencer, 2011, p. 32). Similarly, “images and video open up complex, reflexive and multi-faceted ways of exploring social realities” and provide a uniquely powerful form of ‘thick description’ (Spencer, 2011, p. 34). As qualitative data, video is especially useful for providing ‘thick description’ of a location or program (Oware, Diefes-Dux & Adams, 2007) because it contains and communicates a uniquely large amount of information in an efficient format – hence the cliché that ‘a picture is worth a thousand words.’
Mturk Crowd-Source Platform

The crowd-sourcing aspects of this study were built on the foundation of work by Azzam and Jacobson (2013) related to the potential value of utilizing Amazon’s Mechanical Turk crowd-sourcing platform (henceforth ‘Mturk’) in various elements of evaluation projects. Additionally, this study responds to a question posted on the AEA 365 blog to explore the use of Mturk for qualitative analysis (Azzam, 2013). Readers unfamiliar with Mturk should consult those sources for a helpful introduction and overview.

While the name is unusual and could carry vague suggestions of racism, the meaning behind the name “Mechanical Turk” actually references an interesting historic analogy: it is derived from a mechanical automaton developed at the turn of the 18th century, designed to look like a Turkish “sorcerer,” which could play chess and beat many opponents. While it looked like a technological marvel, the mechanical exterior hid a person working below. In that spirit, Mturk is a web-based platform that seems to ‘automate’ complex tasks, but is in fact facilitating the work of thousands of people that perform ‘human intelligence tasks’ (or ‘HITs’ as they are known on Mturk) such as tagging a photo, taking a survey, or looking up an address (Mason & Suri, 2012). As relates to qualitative research, while there are many programs to facilitate data storage, coding, retrieval, comparing and linking, there is no replacing the need for a human being to do the actual analysis (Patton, 2002, p. 442).

Since its launch in 2005, Mturk has been on a similar path of fast growth like Facebook and YouTube, as it currently has about 500,000 registered workers from 190
countries that are able to select from over 200,000 HITs (Mturk, 2013). Potential employers (known as ‘requesters’) seeking the help of workers set the price they are willing to pay per HIT with a $.01 minimum and seldom over $1.00, and are charged 10% commission for using the service. In order to ensure quality work, workers are paid only after their work is reviewed and approved by requesters, and requesters can set both a minimum number of previously completed HITs (e.g., 500) and an approval rate (e.g., 90%) to address concerns about both experience and quality. While the platform is especially geared to meet the needs of requesters (which are the source of Mturk revenue), other sites address the needs of workers. For example, the mission of Turkopticon is to help “the people in the 'crowd' of crowdsourcing watch out for each other” and protect workers from abusive requesters, which they do by allowing workers to rate requesters on four criteria: “communicativeness, generosity, fairness and promptness” (Turkopticon, 2013).

Studies on Mturk workers found that most spend a day or less per week working on Mturk, and typically earn less than $20 per week (Ipeirotis, 2010). While only 13.8% of the workers in the U.S. reported that Mturk was their primary source of income, 61.4% reported that earning additional money was an important driver of participation (Paolacci, Chandler, & Ipeirotis, 2010). In addition to providing valued but nonetheless supplemental income for most of the workers, 69.6% of the workers in the U.S. reported that they consider Mturk a fruitful way to spend free time, for example a better alternative than watching TV (Paolacci, Chandler, & Ipeirotis, 2010).
Mturk, like crowdsourcing in general, raises a variety of ethical questions, especially related to the potential for ‘digital sweatshop’ conditions of abuse related to low pay or unfair rejection of work (Scholz, 2012, p. vii). While those problems persist, each requester has an opportunity to pay fair wages based on the rate they set, so conscientious evaluators can maintain the ethics of their own practice by paying reasonable rates and maintaining appropriate review procedures. The ethical research principle of privacy and anonymity is effectively built-in to the Mturk platform, as Mturk workers are known only through anonymous worker IDs which do not contain personally identifiable information. This structure is such that many Institutional Review Boards will consider studies ‘exempt’ from full review (Paolacci, Chandler, & Ipeirotis, 2010).

The Mturk platform allows requesters to specify up to 5 criteria for worker selection, including the number of previously completed HITs, approval rate (noted above), geographic requirements, and wide latitude for custom criteria. The ability to geographically target workers ensures that surveys or other tasks requiring cultural-specific knowledge can be targeted appropriately. Toward that point, various demographic studies on Mturk workers in the U.S. indicate that the pool of workers is more demographically diverse than standard Internet samples, and provide significantly more diversity than studies conducted on typical American college campuses (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). While better than many on-campus studies, Mturk workers do not perfectly mirror the US population, as the
overall educational level is higher than the average in the US, and several other demographic categories are overrepresented: younger workers, unmarried, childless, and female (Ipeirotis, 2010).

Increased speed and lower cost are among the primary advantages of using Mturk for research and evaluation projects. One researcher noted that while they “pay $8 for a 15- to 20-minute experiment in a lab…we can run the same study on MTurk for 75 cents to a dollar” (Bohannon, 2011). The Azzam and Jacobson (2013) study paid $.75 per completed survey, and had 500 responses within three days at a total cost (including commission) of $412.50. Because of these advantages, researchers have used Mturk for a wide variety of studies: 10,000 workers developed a tool to track the emotional content of Twitter messages (Bohannon, 2011), deciphering bad handwriting in a process of iterative refinement (Little, Chilton, Goldman, & Miller, 2009), a screening process to identify conscientious Mturk workers (Downs, Holbrook, Sheng, & Cranor, 2010), as a matched comparison group for an evaluation (Azzam & Jacobson, 2013), and numerous conducting surveys (e.g., Paolacci, Chandler, & Ipeirotis, 2010; Ipeirotis, 2010) just to illustrate the range.

The quality of data using Mturk has also been addressed in a number of studies, with a growing number indicating that results can be comparable or even superior to established alternatives. While collecting data through the Internet has clear limitations, results have been shown to be consistent with traditional approaches and provide better diversity as compared with traditional samples (Gosling, Vazire, Srivastava, & John, 2004). Test-retest reliabilities were found to compare favorably
with correlations of traditional methods, and the data provided by MTurk “met or exceeded the psychometric standards associated with published research” (Buhrmester, Kwang, & Gosling, 2011, p. 5). Similarly, Mturk workers annotating affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation were found to be in high agreement with experts (Snow et al., 2008). One researcher summarized this growing body of research by saying, “there are numerous studies that show correspondence between the behavior of workers on Mechanical Turk and behavior offline or in other online contexts…evidence that Mechanical Turk is a valid means of collecting data is consistent and continues to accumulate” (Mason & Suri, 2012, p. 4).

The Study

“The Medieval alchemy aimed to transmute base metals into gold. Modern alchemy aims to transform raw data into knowledge, the coin of the information age…Fine qualitative analysis remains rare and difficult - and therefore valuable” (Patton, 2002, p. 432). While this study is conducted within the very broad context of user-generated visual data and crowd-sourcing, the focal research question guiding this study is specific: How does crowd-sourced open coding of video data compare with coding provided by a control group of trained evaluators? This question is inspired by the goal of developing a new methodology to contribute to the difficult process of ‘transforming raw data into knowledge’ with the speed and emotional relevance
needed to contribute to the feedback mechanisms at the heart of complexity. This study is specifically addressed toward crowd-sourcing the “open coding” (Corbin & Strauss, 2008) step of the analysis, part of the initial inductive phase of analysis where “findings emerge out of the data…in contrast to deductive analysis where the data are analyzed according to an existing framework” (Patton, 2002, p. 453). As described by Strauss & Corbin (2008), ‘open coding’ is a process that:

“…requires a brainstorming approach to analysis because, in the beginning, analysts want to open up the data to all potentials and possibilities contained within them. Only after considering all the possible meanings and examining the context carefully is the researcher ready to put interpretative conceptual labels on the data. Conceptualizing the data not only reduces the amount of data the researcher has to work with, but at the same time provides a language for talking about the data.” (p. 160)

The present study uses video data in order to demonstrate the analytic potential for this form of communication which is typically relegated only to illustration or documentary purposes (Emmison & Smith, 2000, p. ix), and builds on others which utilized crowd-sourced workers for analysis (e.g., Bohannon, 2011; Little, Chilton, Goldman, & Miller, 2009; Downs, Holbrook, Sheng, & Cranor, 2010). While the proposed crowd-sourced analysis methodology is unique, elements are related to internet ethnography, which conceptualizes the internet as a “communication medium, a global network of connection, and a scene of social construction… (providing) new tools for conducting research, new venues for social research, and new means for
understanding the way social realities get constructed and reproduced through discursive behaviors” (Markham, 2004, p. 95). “This medium is seen as both a tool and a site for qualitative research, developed from the observation that social life in contemporary society communicates, interacts, and lives more online; for ethnographers to better understand the ‘social world,’ they must adjust their research methods to reflect these changes” (Garcia et al. 2009).

Method

The method for this study was developed to address the focal research question, and is presented in the five categories of design, instrumentation, procedure, sample, and analytic approach.

Design

The design of the current study was developed to complement and sequentially precede a companion study (Wilson & Azzam, manuscript in preparation), which explored the use of crowd-sourced participants in the ‘categorization coding’ stage of analyzing video data. While both studies utilize Mturk and the same video data, they address distinctly different questions and stages of the coding process. The research questions of the current study led to a quasi-experimental design that compared the coding decisions provided by ‘treatment’ group of crowd-sourced participants with those provided by the ‘control’ group of trained evaluators.
To facilitate this geographically diverse group of participants, the video coding was operationalized into internet surveys hosted on SurveyMonkey.

**Instrumentation**

The first group of coding prompts on the internet survey asked participants to “enter 3-5 descriptive words about the content during each 30 seconds of video,” followed by a series of short answer boxes for each 30 second segment of time. Participants were allowed the flexibility to choose the number of descriptions they provided in order to compare the quantity of coding data provided by crowd-sourced and trained evaluator participants and test potential variance in quality of responses.

The coding prompt following the 30 second increments asked participants to “please provide a narrative description of the video & the impact it had on you - as if you were telling a friend” and large textbox for a descriptive paragraph in order to compare this time-specific coding with open-ended description. These two structures of coding prompts were followed by a second page of the survey which contained a series of demographic questions (e.g., age, gender, education) and a ‘survey completion code’ which the Mturk workers cut/pasted into a box on the Mturk platform to get credit for completing the task on the SurveyMonkey platform. These surveys were pilot-tested with 5 Mturk workers each, and following this test the initial data was analyzed and minor refinements to the prompts were completed to clarify the task and structure.
**Procedure**

The first step in this study was to create a ‘requester’ account on Mturk.com and complete a series of small pilot tests of crowd-sourced surveys and video coding to learn more about the platform and develop familiarity with the platform. While Mturk provides an HTML based interface with a selection of templates for common tasks or HITs, as well as freedom for extensive customization through a Java-based ‘application programming interface,’ neither were used for this project because of the additional technical needs and complications. Instead of developing the survey directly on the Mturk platform, one web-based survey per video (for a total of 5) was developed on SurveyMonkey.com. When participants clicked on the link within Mturk, they were directed to the SurveyMonkey website containing the survey which included the product review video as an embedded file to minimize barriers to use. Following a pilot test and minor text revisions, the survey was distributed to participants through either the Mturk platform or emailed to the evaluator control group.

**Sample**

The survey was distributed to 25 Mturk participants per video survey for a total of 125 participants. The worker criteria on the Mturk platform included that they be located in the United States, have completed at least 500 approved HITs with an approval rating of 90%, and were paid between $.50 and $.75 depending on the length of the video.
The selection and composition of the control group was intended to be similar to existing qualitative research methodologies, such that the participants were recruited based on their experience, training, and availability in their schedule. The control group participants included the primary researcher plus two fellow students in the Western Michigan University Interdisciplinary PhD in Evaluation program. The trained evaluators serving as the control were not recruited or paid through Mturk, so the SurveyMonkey links were emailed to them directly and they were each offered a $20 gift card as appreciation for their more extensive involvement.

The survey link was available to all Mturk workers meeting the basic criteria, and the final sample was based on factors such as being on the platform while the link was live and participant selection of this task. The demographic characteristics of the Mturk workers which participated in this study were broadly consistent with the findings from studies of Mturk workers (e.g., Ipeirotis, 2010). An analysis of IP addresses available through SurveyMonkey indicated that 87 individuals completed one survey, 15 completed two and 4 completed 3, for a total of 106 unique Mturk participants. That said, in order to maintain consistency with the findings, the participant demographics presented in Table 1 were compiled based on the total 129 Mturk surveys.
### Table 5

*Participant Demographics*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Category</th>
<th>Mturk</th>
<th>Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>39%</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>61%</td>
<td>67%</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18-29</td>
<td>49%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>30-49</td>
<td>38%</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>50 and over</td>
<td>13%</td>
<td>33%</td>
</tr>
<tr>
<td><strong>Employment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Employed 40 hours/week</td>
<td>38%</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>Employed 1-39 hours/week</td>
<td>35%</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>Unemployed, looking for work</td>
<td>14%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Unemployed, retired or disabled</td>
<td>13%</td>
<td>-</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High school graduate</td>
<td>16%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Trade/technical/other</td>
<td>4%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Some college</td>
<td>40%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>College graduate</td>
<td>33%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Post graduate degree</td>
<td>8%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Marital</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5 - continued

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Category</th>
<th>Mturk</th>
<th>Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single/never been married</td>
<td>62%</td>
<td>33%</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>30%</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>5%</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Separated</td>
<td>2%</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Race**

<table>
<thead>
<tr>
<th>Race</th>
<th>Mturk</th>
<th>Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>69%</td>
<td>100%</td>
</tr>
<tr>
<td>Asian</td>
<td>12%</td>
<td>-</td>
</tr>
<tr>
<td>Black or African-American</td>
<td>8%</td>
<td>-</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>6%</td>
<td>-</td>
</tr>
<tr>
<td>White Asian, Other</td>
<td>5%</td>
<td>-</td>
</tr>
</tbody>
</table>

In addition to the demographics above, Mturk participants came from a total of 28 states representing all regions in the lower 48 states; 24 states had 2 or more participants each, while California had the most participants (21% of total) with Texas second with 7% of the total. As context for the racial diversity, all census categories except American Indian and Native Hawaiian were represented, but Black/African Americans are slightly underrepresented and Asians slightly overrepresented as compared with demographic ratios from a recent census report (Martin et al., 2010).

The selection of video data to serve as the basis for coding consisted of five user-generated product review videos hosted on the YouTube channel of
ExpoTV.com, which is “an online community of consumers who share their unbiased, honest opinions in video” (ExpoTV, 2013). These videos were intentionally sampled to maximize the diversity of product types, demographics of the presenters, lengths, and the popularity of the videos as indicated by the number of views it had received. An overview of the five videos is provided in Table 2:

Table 6

*User-Generated Product Review Videos* Coded by Study Participants

<table>
<thead>
<tr>
<th>Video #</th>
<th>Product</th>
<th>Length</th>
<th>Ethnicity</th>
<th>Gender</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Samsung A900 Phone</td>
<td>1:08</td>
<td>White</td>
<td>Female</td>
<td>28,290</td>
</tr>
<tr>
<td>2</td>
<td>Apple 8 GB iPod</td>
<td>2:32</td>
<td>White</td>
<td>Male</td>
<td>24,501</td>
</tr>
<tr>
<td>3</td>
<td>Settlers of Catan Game</td>
<td>1:42</td>
<td>White</td>
<td>Female</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>St. Ives Facial Moisturizer</td>
<td>3:03</td>
<td>Af. Am.</td>
<td>Female</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>Toy Helicopter</td>
<td>2:30</td>
<td>Asian</td>
<td>Male</td>
<td>2</td>
</tr>
</tbody>
</table>

The primary data for this study is the descriptive codes provided by both the Mturk workers and the three evaluators as a control group for comparison purposes. While 25 participants per survey were paid through the Mturk platform, the SurveyMonkey data showed that surveys 1 and 5 each had one extra completed and survey 2 had two extra completed, for 129 surveys completed by Mturk participants plus 15 by the three evaluators for a total of 144 completed surveys. The likely reason for the extra Mturk surveys is that they could be completed on SurveyMonkey without

---

1 Citations for the video data are marked with an ‘*’ in the References
participants taking the final step of entering the unique ‘completion code’ on the
Mturk platform; participants that forget that step would not be paid nor counted on
Mturk, resulting in additional individual(s) being linked to the survey.

Analytic Approach

The qualitative analysis conducted for this study is within the broad category
of “content analysis,” a “qualitative data reduction and sense-making effort that takes a
volume of qualitative material and attempts to identify core consistencies and
meanings” (Patton, 2002, p. 453). As such, the prompts for the internet surveys
included the language to “describe the content” in both 30 increments and as an
overall descriptive paragraph. Consistent with the content analysis approach, the
analysis of the analytic data provided by study participants includes both word counts
and analysis of broad categories or themes (Marshall & Rossman, 2011, p. 161). It
should be noted that the process of analyzing the data did not begin with a
‘preordained operational variable’ imposed by the researcher, with the intent of
avoiding the potential bias associated with “imposing a limited worldview” (Marshall
& Rossman, 2011, p. 91) on the findings.

Findings

Two primary comparisons highlight the potential strengths and weaknesses of
this methodology: the comparison of more detailed coding of the video data in 30
second increments with a more holistic descriptive paragraph, and the comparison of
the coding completed by the Mturk participants with that provided by evaluators.
Table 3 presents the average word counts for each video within each comparison category as well as an overall average. Additionally, while not evident in the overall word count data in the table, it should be noted that the Mturk participants completed 81% of the total survey fields for the 30 second coding, well more than both the 69% completed by the evaluators and the 60% that was required based on the study design and prompt (i.e., 3 of the 5 cells per 30 second increment).

An additional element of context for the word count comparison is that the variation in word count between individual videos in the 30 second coding for Mturk participants directly corresponds with the length of the videos since longer videos require more coding, and vary from a low of 31 words for the 1:08 video to a high of 72 for the 3:03 for an overall average of one word every 2 seconds. The considerably different pattern for the word count provided by the evaluator comparison group is discussed in the section below - ‘Comparing Mturk Participants with Evaluators.’

Table 7

Word Count Comparison

<table>
<thead>
<tr>
<th></th>
<th>30 Second Coding</th>
<th>Descriptive Paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mturk</td>
<td>Evaluator</td>
</tr>
<tr>
<td>Video 1</td>
<td>31</td>
<td>52</td>
</tr>
<tr>
<td>Video 2</td>
<td>58</td>
<td>114</td>
</tr>
<tr>
<td>Video 3</td>
<td>44</td>
<td>57</td>
</tr>
<tr>
<td>Video 4</td>
<td>72</td>
<td>98</td>
</tr>
<tr>
<td>Video 5</td>
<td>67</td>
<td>56</td>
</tr>
</tbody>
</table>
A qualitative content analysis of the Mturk and evaluator coding, presented in Table 4, revealed that the content could be summarized within four primary categories or themes, including: product ID (e.g., ‘Apple iPod touch’), presenter description (e.g., ‘she slurred her words’), product description (e.g., ‘10 oz. jar’) and overall impact (e.g., ‘very convincing…I would definitely buy it’). Within each of these broad categories the data varied across related dimensions (e.g., full, none or partial product description) and Table 4 presents the relative weighting of the codes used by the Mturk and Evaluator groups. The percentages for each category and coder type total 100% (adjusting for rounding) to both equalize the different size comparison groups (25 for Mturk and 3 for Evaluator) and highlight the primary qualitative differences.
### Table 8

*Mturk and Evaluator Qualitative Comparison*

<table>
<thead>
<tr>
<th>Category</th>
<th>30 Second Coding</th>
<th>Descriptive Paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mturk</td>
<td>Evaluators</td>
</tr>
<tr>
<td><strong>Product ID</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>39%</td>
<td>27%</td>
</tr>
<tr>
<td>None</td>
<td>44%</td>
<td>73%</td>
</tr>
<tr>
<td>Partial</td>
<td>17%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Presenter Description</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>15%</td>
<td>53%</td>
</tr>
<tr>
<td>Neutral</td>
<td>18%</td>
<td>20%</td>
</tr>
<tr>
<td>None</td>
<td>51%</td>
<td>20%</td>
</tr>
<tr>
<td>Positive</td>
<td>16%</td>
<td>7%</td>
</tr>
<tr>
<td><strong>Product Description</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed</td>
<td>49%</td>
<td>53%</td>
</tr>
<tr>
<td>Moderate</td>
<td>30%</td>
<td>33%</td>
</tr>
<tr>
<td>None</td>
<td>22%</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Impact</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>19%</td>
<td>7%</td>
</tr>
<tr>
<td>Neutral</td>
<td>2%</td>
<td>47%</td>
</tr>
<tr>
<td>None</td>
<td>66%</td>
<td>27%</td>
</tr>
<tr>
<td>Positive</td>
<td>14%</td>
<td>20%</td>
</tr>
</tbody>
</table>
Comparing Incremental Coding with Descriptive Paragraphs

The overall average word count for the two types of analytic data is roughly similar for Mturk participants – 54 for 30 second coding vs. 49 for overall descriptive paragraphs. Though the quantity of data was similar, the analysis revealed qualitative differences between the two types of data. When providing coding in 30 second increments, the Mturk participants were more likely to say nothing about either the product ID (44% vs. 33%), description of the presenter (51% vs. 36%), nor the impact the video had on them (66% vs. 34%). Instead, Mturk participants were much more likely to provide detailed product description when providing coding in 30 second increments as compared with a descriptive paragraph (49% vs. 23%).

The evaluator group had a different profile related to the different types of data. The average word count was substantially higher for 30 second coding than the descriptive paragraphs (75 vs. 49). While the evaluators were equally unlikely to provide a product ID (73% vs. 73%), they were more likely to describe the presenter in the 30 second coding than the descriptive paragraphs (80% vs. 60%).

This pattern of response from both Mturk participants and evaluators is intuitive: Coding in 30 second increments elicited more analytic data (higher overall word count) and was much more likely to elicit responses related to explicitly stated or visual features of the video (i.e., the product being described) and less likely to elicit coding related to intangible, emotional impact aspects of the video data. Another overall finding is that the majority of Mturk participants took the descriptive coding
tasks seriously and provided thoughtful, useful analytic data; even the few that skipped most of the 30 second coding cells or provided crude or otherwise unprofessional responses (described further below) still provided a perspective that is found within elements of the American public.

Comparing Mturk Participants with Evaluators

One goal of the early stages of qualitative analysis is to “define relevant variables of representation and/or salience” (Bell, 2001, p. 15) and a primary overall finding from this comparison is that there are marked differences between trained evaluators and Mturk participants in the features of the video data which are considered ‘relevant’ for coding as well as in the total volume of coding words provided.

The overall average word count of evaluators was higher for 30 second coding (75 vs. 54) but identical for the descriptive paragraphs (49 each). These similarities disappear, however, when considering the length of the videos. The word count variation of 30 second coding for Mturk participants directly corresponded with the length of the video with a per-video average of roughly 1 word every two seconds. This contrasted with the time-adjusted word count for evaluators, which decreased for each video 1 through 5; the high for video 1 was 1 with a word every 1.3 seconds and the low was video 5 with a word every 2.7 seconds. This trend was even more pronounced for the descriptive paragraph, with the evaluators providing one word every 1.1 seconds for video 1 to one word every 6.3 seconds for video 5.
The qualitative differences between the two groups were equally pronounced. Evaluators were substantially less likely to provide a product ID in either the 30 second coding or descriptive paragraph (73% ‘none’ vs. 44% and 33%) and were unlikely to provide a product description in the descriptive paragraph (73% vs. 49%), but more likely to describe the presenter in negative terms (53% vs. 15%) and express an overall neutral impact in both the 30 second coding and descriptive paragraph (47% for both forms of data vs. 2% and 11% for Mturk participants). On a more specific level, a few of the Mturk participants responded to the tight shirt of the blond woman in video 1 with very unprofessional, sexual descriptions, and many were considerably more ‘raw’ than the evaluators, indicating (for better or worse) that they felt considerable freedom in the anonymity provided through Mturk.

These differences suggest a few overall conclusions. The overall similarity in average word counts and high proportion of total cells filled by Mturk participants (81%) was well beyond the 60% required by the prompts indicates that most of the Mturk workers took their job seriously and provided a good volume of analytic data. The consistent average word count for Mturk participants as compared with the decreasing average word count for the evaluator comparison group suggests the potential value Mturk provides through ‘fresh’ analytic perspective; the comparison group (similar to most traditional qualitative analysis teams) reviewed each of the 5 videos, while only a few Mturk participants reviewed more than one. One interpretation is that the decreasing word count for the evaluators is related to either fatigue or task familiarity – neither of which were evidenced by the Mturk
participants. The overall qualitative differences between Mturk participants and evaluators are also intuitive, as the coding and descriptions provided by the evaluators was more professional and deliberately ‘evaluative’ than the more casual (sometimes crass) descriptive data provided by Mturk participants.

The qualitative and quantitative coding differences are related to differences between the Mturk participants and the evaluator control group in underlying values, training and conceptual framework. While there are some projects where the evaluator perspective or conceptual frame is explicitly needed, in many evaluation and qualitative research projects the bias associated with researcher perspective is a threat or conceptual barrier to reflecting the values of the participants or a broader group. For example, in the tradition of phenomenology, this threat is addressed through the process of ‘bracketing’ which is when “investigators set aside their experiences, as much as possible, to take a fresh perspective toward the phenomenon under investigation” (Creswell, 2007, p. 59-60). This finding suggests that utilizing Mturk participants in the coding is an alternate, and perhaps more direct, means to accessing a fresh perspective on the video data.

Discussion

The findings address the focal research question about how crowd-sourced open coding of video data compares with coding provided by a very small and homogeneous control group of trained evaluators based on several comparisons. Deeper consideration of these comparisons reveals that this methodology offers a
number of potential benefits, especially within contexts of complexity, but that these benefits must be approached with caution and awareness of several important limitations.

*Potential Benefits*

Crowd-sourcing has the potential to be a cost and time effective way to accomplish a variety of tasks, and the findings of this exploratory study provided initial indications that there are various potential benefits available with crowd-sourcing video data analysis. As relates time, this study did require a time investment for the initial development of this approach, including setting up the accounts, developing the online surveys with embedded videos, etcetera. However, once those initial steps were completed, the analytic data from Mturk was available within 12 hours, as compared to the week between the initial request and the receipt of data from the evaluation team. As for cost, the 129 surveys cost a total of $74.52 (Mturk participants plus Amazon commissions) for an average of $.596 per completed survey, very competitive with other methods using small teams and well below the cost of accessing analysts of similar diversity.

In addition to time and cost benefits, the primary benefit of this methodology is that it addresses the longstanding principle of qualitative research which is that “the whole analysis of experience must be based on their concepts, not ours” (Boas, 1943, p. 314), or for the analyst to get “out of the way of the data to let the data tell their own story” (Patton, 2002, p. 457). Clear as this goal is, the challenge has always been the
practical means to accomplish it, and eliciting analytic data from the ‘crowd’ through Mturk provides an interesting new opportunity toward that end. By receiving the analytic perspective of a wider audience, the methodology contributes to the analytic process of ‘Epocche’ which is a “process that the researcher engages in to remove, or at least become aware of, prejudices, viewpoints, or assumptions regarding the phenomenon under investigation” (Patton, 2002, p. 485). This methodology is consistent with many qualitative research approaches which seek for ways that “looking precede judgment and that judgment of what is ‘real’ or ‘most real’ be suspended until all the evidence (or at least sufficient evidence) is in” (Ihde, 1986, p. 36).

Furthermore, the broad geographic and demographic diversity of Mturk workers is such that selection procedures could be established to recruit representatives (or proxies) of program participants in many evaluation contexts (e.g., adults in a specific state, women in a region of the country). With Mturk participants providing representative feedback on the early analysis process, their insights could help shape the analytic frame in the formative process, a potentially significant supplement (but not replacement) to the current practice of a post-analysis ‘member check’ (Patton, 2002; Seale, 2002) of the findings after the analytic frame is established. In this way, the Mturk platform mediates the communication between the non-expert analysts and the evaluator in ways that shift the power dynamic to empower the voice and perspective of the non-expert participants (Markham, 2004; Garcia et al., 2009). Additionally, including participant representatives in the analytic
process could contribute to or supplement some of the goals of collaborative and participatory approaches (Patton, 2002, p. 496).

In qualitative studies “the researcher is the instrument,” (Marshall & Rossman, 2011, p. 112) such that the “human factor is the great strength and the fundamental weakness of qualitative inquiry and analysis – a scientific two-edged sword” (Patton, 2002, p. 433). This methodology is still dependent on the human factor, but by employing a crowd instead of a single researcher or small research team, offers a new opportunity to address the limitation of individual bias by replacing it with the strength (and different set of limitations) related to diverse perspectives. For example, this study revealed how the coding provided by the 3 evaluators was clearly influenced by their perspective (highly educated, professional, evaluative) while the varied perspectives of Mturk participants reflected differing levels of education and values; depending on the project goals, this difference in perspective could be either a useful benefit or barrier and reason to utilize existing methodologies.

**Benefits in Context of Complexity**

This study was conducted within the context of “the great unexplored frontier (of) evaluation under conditions of complexity” (Patton, 2011, p. 1). While there are many understandings of the term, for the purposes of this discussion, ‘complexity’ is “characterized by a large number of interacting and interdependent elements” (Mitchell, 2009, p. 13) which operate in a nonlinear system webbed with feedback loops that amplify or otherwise inform ongoing growth and change (Wheatley, 1999).
While evaluators seek clarity and warrantable evaluative claims, these interrelationships and feedback loops seriously undermine their efforts, leading some to conclude that complexity is “the most profound and limiting methodological issue” (Pawson, Wong, & Owen, 2011, p. 523).

The inherent inter-relationships in complex programs and contexts make it exceedingly difficult to reduce the issues involved to their component parts. Patton has suggested an approach called ‘double loop learning’ as a way to influence and understand the complexities. This approach makes “changes to the system either to prevent the problem or to embed the solution in a changed system” (Patton, 2011, p. 11). By informing or influencing the feedback mechanisms at the heart of complex systems, evaluators can harness the mechanisms of complexity toward improved problem solving, not simply providing detached measurements or reports but rather serving as active agents of change.

Limitations

This study utilized user-generated product review videos as the primary source of data, so is delimited within that context. As a small scale pilot-study, this study is also inherently limited to this single point in time, by the small dataset, and relatively small group of participants. Additionally, the methodology was also limited by the fact that neither the selection of Mturk participants nor comparison sample was random. That said, this selection bias was not considered a threat to validity for this study: Mturk HITS commonly require similar qualifications utilized in this study so that the
opportunity to participate was considered as broad as typical for Mturk. All qualified workers were equally free to login and access the surveys when they were posted, so while this was not random, it was an unguided participant selection process within the timeframe the links were active. Similarly, the participant selection of the evaluator control group was based on common eligibility practices, so while not representative of evaluators or qualitative researchers in general, this was also not considered a threat to validity. The different size of the groups in the comparison is another limitation of the methodology (3 in control vs. 25 Mturk participants per survey) but this factor is a basis for the comparison under study, as qualitative analysis coding is typically completed by 1-3 evaluators because of cost and time constraints not operative using the Mturk platform. The difference in group size was addressed by calculating overall group averages and conducting analysis at the group level, not sums of individual responses. In conclusion, while this study methodology did involve a range of limitations, none were considered serious threats to the validity of the overall findings.

**Future Research and Applications of Methodology**

An exploratory companion study seeking to build on the findings of this study was conducted (Wilson & Azzam, manuscript in preparation) which utilizes Mturk participants for coding video data within categories developed from the open-code data of the present study. The intent of linking the two studies by using the same conceptual frame and data was to explore the potential of Mturk coding in two sequential stages of a qualitative research project. Beyond that study, several avenues
for future research are related to expanding the breadth and depth of this exploratory study and addressing the limitations noted above. For example, other forms of video data (e.g., video recorded interviews, classroom observations, training sessions, etc.) could be analyzed by Mturk participants. Similarly, researching Mturk coding of larger datasets and/or studies conducted over longer periods of time would add depth and needed nuance to the methodology. Finally, coding other forms of qualitative data (e.g., audio recordings, still photos, text) should also be studied, as the analysis process for these other forms of qualitative data retain important differences that should be explored utilizing crowd-sourced analysis.

As an exploratory study, the findings suggest several initial directions for evaluators to consider for application of this method, including:

1. The perspectives provided in the analytic data (in detailed coding such as the 30 second increments, broader descriptive paragraphs, or both) could be analyzed to guide the development of a coding and analysis scheme that is based on a wider viewpoint than the evaluator or evaluation team.

2. Mturk participants could provide valuable analytic data in the initial overview stage of analyzing video data. Both the speed of feedback and additional perspectives could be useful in the process of trying to “Get a sense of the whole” (Patton, 2002, p. 440, emphasis in the original) of a large body of video data – or potentially other qualitative data. If the body of data is very large, excerpts or clips could be randomly sampled to provide a representative sample of data for Mturk participant review.
3. The Mturk workers could be pre-qualified for participation based on a demographic survey and selected based on geographic or demographic parameters similar to program participants, and as such could be treated as representatives of program participants, providing even greater utility and conceptual benefit from the data and perspective they provide.

4. The general description of video data such as the descriptive paragraphs could be compiled to a more broad and nuanced “thick description…the foundation for qualitative analysis and reporting” (Patton, 2002, p. 437) than possible with the limited perspective of an evaluator or small team.
III. TOWARD A QUALITATIVE ANALYSIS DREAM TEAM: EXPLORING CROWD-SOURCED VIDEO DATA CODING

Authors

Kurt Wilson, Western Michigan University, Tarek Azzam, Claremont Graduate University, Chris L. S., Coryn, Western Michigan University and Daniela Schroeter, Western Michigan University

Abstract

This comparative study explores a new methodology for categorical coding of video data utilizing crowd-sourced participants. An overview of the history, strengths, and limitations of the Mturk crowd-sourcing platform is provided as background, and a range of studies utilizing crowd-sourcing for research and evaluation are discussed. The video data coded are user-generated reviews of various products, and coding is completed by cohorts of 3, 10, and 20 coders which represent considerable demographic and geographic diversity.

The findings indicate that the coding provided by crowd-sourced participants is nearly perfectly agreement for the gender of the presenter, is generally high for objective details presented in the video content, and shows mixed degrees of agreement for a range of other coding decisions. The broad conclusions from this study suggest that coding of video data can be completed in a matter of hours at a reasonable cost, and that this methodology potentially reduces the barriers to use for video data and offers intriguing new applications.
Keywords: Video analysis, Mturk, qualitative analysis, crowd-sourced, user-generated data, product review video, ExpoTV

Introduction

Still and moving images are forms of communication with a long history: photography was invented in 1798 (Gernsheim, 1986) and Thomas Edison patented the first movie camera in 1888 (Green, 2013). More recent advances in digital imaging and social media built on this foundation and accelerate the established trends of progressively lowering costs and broadening usage. Low-cost digital photography available in cellphones marked a particularly significant milestone in 2010, as over one billion camera phones were in use worldwide by 2010 (The Economist, 2010). As just one indicator for the total volume of photos currently captured, 300 million photos are uploaded to Facebook each day, and the total photo archive on Facebook currently exceeds 240 billion photos (Madrigal, 2013). In 2008, users were viewing photos on Facebook at a rate of 300,000 per second. Moreover, conservative estimates of current consumption based on underlying growth trends indicate that several million photos are viewed each second on Facebook (Madrigal, 2013). As context for those substantial numbers, Facebook was launched as a student project less than 10 years ago (Tabak, 2004), indicating both the speed of cultural change and broad underlying demand for this type of visual communication. As for video, YouTube was launched in 2007 and currently has over 1 billion unique visitors watching 6 billion hours of
video each month, and new video content is uploaded at a rate of over 100 hours of footage every minute (YouTube, 2013). Within the context of the total visual culture, the ‘user-generated’ visual communications on Facebook and YouTube are a small part of the whole, as the full visual environment includes professional film productions, advertisements, magazines, newspapers, and television.

These examples indicate that visual communications are a substantial element of global culture (especially in the developed world) and an important aspect of current communication practices. Given the growing concern of cultural competence for evaluators (American Evaluation Association, 2009) and the potential richness of visual data, serious consideration of visual data is a requirement of professionalism with considerable promise. Sadly, there are only a few examples of visual evaluation methods presented in the evaluation literature of the past thirty years (e.g., Ingle, 1984; Fang, 1985; Bennington, Gay, & Jones, 1999; Tucker, & Dempsey, 1991; Hurworth, & Sweeney, 1995; Bessell, Deese & Medina, 2007; Oware, Diefes-Dux, & Adams, 2007). In the broader research community, visual methods are also relatively scarce, but nonetheless are an established specialty (Banks, 2001) represented by both the International Visual Sociology Association (IVSA) and Commission on Visual Anthropology (CVA), each of which host their own conferences and journals, Visual Studies (IVSA, 2013) and Visual Anthropology (CVA, 2010), respectively.
User-Generated Video Data

While no published examples of user-generated video data used in program evaluations exist, the many product review videos and patient testimonials posted on YouTube do provide relevant examples from the evaluation enterprise more broadly defined. Furthermore, evaluative videos of this nature are themselves part of the growing body of crowd-sourced evaluation data, which includes Amazon or Zagat reviews. Michael Scriven referred to this form of data in his brief paper “Ultimate Evaluation Questions,” and said it “should be treated as a valuable—indeed often invaluable—source of information” (Scriven, 2013, p. 3-4).

Visual data, and video in particular, are an especially powerful form of communication. The “immediacy of the visual affects us in a profound and elusive way—before the sense-making apparatus, the cognitive processing, there is a pre-reflective reaction, as several writers and researchers have noted. There seems to be agreement that there is something indefinable about the visual, grounding it in material reality. It is an immediate and authentic form which verbal accounts are unable to fully encompass” (Spencer, 2011, p. 32). Similarly, “images and video open up complex, reflexive and multi-faceted ways of exploring social realities” and provide a uniquely powerful form of ‘thick description’” (Spencer, 2011, p. 34). As qualitative data, video is especially useful for providing ‘thick description’ of a location or program (Oware, Diefes-Dux & Adams, 2007) because it contains and communicates a uniquely large amount of information in an efficient format—hence, the cliché that ‘a picture is worth a thousand words.’ Additionally, video can be a particularly effective medium for
teaching or institutional learning efforts, as “watching ‘real’ people from different communities talk about their lived experiences is refreshing to students because it captures an embodied expression, not abstract truisms…” (Spencer, 2011, p. 33).

**Video Data Analysis**

A distinct analytic advantage of visual data in general is that it “is closely analogous to perceiving reality because photographs provide a point-by-point correspondence to what was in front of the camera...” (Van Leeuwen, 2001, p. 94). Researcher controlled factors such as framing, lighting, and focus are certainly critical considerations with associated limitations but, in general, visual data provide a substantially more nuanced and faithful record than possible through words or numbers. However, this great strength of data richness is also a great weakness, as “it can then be difficult to codify and categorize video data, to break it apart into fragments that can be subjected to analysis, and transcriptions of visual conduct can seem unwieldy and time-consuming to produce, even for the shortest fragment” (Luff & Heath, 2012, p. 258). Most commonly this difficulty is addressed through focusing on words as the basic form in which the data are found” (Miles & Huberman, 2013, p. 71) which for visual data requires an initial step of writing memos about photographs or transcribing video content into text, both of which are “fraught with slippage” (Miles & Huberman, 2013, p. 71).

Theoretical assumptions or evaluation goals form the foundation of a typical video analysis project, and these assumptions or goals are then operationalized into
categories (Knoblauch & Schnettler, 2012). The primary difficulties include subjectivity in formulating the categories, lack of reliability in coding or categorizing, and the extensive time required for coding and analysis (Walberg, Lu, Niemiec, & Walberg, 1997).

To categorize, or code, such information, “data are broken down, compared, and then placed into discrete categories. Similar data are placed in similar categories, and different data create new categories. “Coding is an iterative, inductive, yet reductive process that organizes data, from which the researcher can then construct themes, essences, descriptions, and theories” (Walker & Myrick, 2006, p. 549).

**Mturk Crowd-Source Platform**

The crowd-sourcing platform provided by Amazon is called the “Mechanical Turk” (henceforth ‘Mturk’). While this name is unusual and could convey a vague suggestion of racism, the meaning references an interesting historic analogy: it is derived from a mechanical automaton developed at the turn of the 18th century which was designed to look like a Turkish “sorcerer” and could play chess (Mason & Suri, 2012). While to the general public it seemed like a technological marvel, the mechanical exterior actually hid a person working below, which was the true source of the genius. In that spirit, Mturk appears to ‘automate’ complex tasks, but is in fact a platform for facilitating the work of thousands of people that perform ‘human intelligence tasks’ (or ‘HITs’ as they are known on Mturk) such as tagging a photo, taking a survey, or looking up an address (Mturk, 2013, paragraph 1). As relates to
qualitative research, while there are many programs to facilitate data storage, coding, retrieval, comparing and linking, there is no replacing the need for a human being to do the actual analysis (Patton, 2002, p. 442), such that the human involvement facilitated by Mturk is critical.

Since its launch in 2005, Mturk has experienced exponential growth and currently has about 500,000 registered workers from 190 countries that are able to select from over 200,000 HITs (Mturk, 2013, paragraph 3). Potential employers are referred to as ‘requesters,’ and are able to set the price they are willing to pay per HIT ($0.01 minimum and seldom over $1.00), and an additional 10% commission for Mturk is added to their final bill. In an effort to ensure high quality work, workers are paid only after their work is reviewed and approved by requesters, and requesters can set both a minimum number of previously completed HITs (e.g., 500) and an approval rate (e.g., 90%) to address concerns about both quality and experience. While the platform is especially geared to meet the needs of requesters (which provide Mturk commission revenue), other sites address the needs of workers. For example, the mission of Turkopticon is to help “the people in the 'crowd' of crowdsourcing watch out for each other” and protect workers from abusive requesters, which they do by allowing workers to rate requesters on four criteria: “communicativeness, generosity, fairness and promptness” (Turkopticon, 2013, paragraphs 1 and 4).

Studies of Mturk workers have found that most workers spend a day or less per week working on Mturk, and typically earn less than $20 per week (Ipeirotis, 2010). While only 13.8% of workers in the U.S. reported that Mturk was their primary source
of income, 61.4% reported that earning additional money was an important driver of participation (Paolacci, Chandler, & Ipeirotis, 2010, p. 412). The income is therefore a valued but supplemental income for most of the workers, and about 70% of the workers in the U.S. report the additional motivation that Mturk provides a fruitful way to spend free time—a better alternative than sitting around or watching TV (Paolacci, Chandler, & Ipeirotis, 2010, p. 413).

Mturk, like crowdsourcing in general, is “… a relatively new ethical and legal territory, and therefore the policies surrounding them are open to debate” (Mason & Suri, 2012, Ethics and Privacy section, para. 2). Among the various ethical questions raised is the potential for ‘digital sweatshop’ conditions related to low pay or unfair rejection of work (Scholz, 2012, p. vii). While these risks persist, a partial solution is built into the structure, as each requester (i.e., evaluator or researcher using the service) can readily maintain the ethics of their own practice by paying reasonable rates and maintaining appropriate review procedures. The ethical research principle of privacy and anonymity is also effectively built-in to the Mturk platform, as Mturk workers are known only through anonymous worker IDs, which do not contain personally identifiable information. This structure is such that many Institutional Review Boards (IRBs) will consider such studies as ‘exempt’ from full IRB review (Paolacci, Chandler, & Ipeirotis, 2010).

The Mturk platform also allows requesters to specify up to 5 criteria for worker selection, including the number of previously completed HITs, approval rate, geographic requirements, and broad freedom to develop custom criteria. The ability to
geographically target workers ensures that surveys or other tasks requiring culturally-
specific knowledge can be appropriately targeted. That being said, various
demographic studies on Mturk workers in the U.S. indicate that the pool of workers is
more demographically diverse than standard Internet samples, and provide
significantly more diversity than studies conducted on typical American college
campuses (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011;
Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010).

Some of the primary advantages of using Mturk for research and evaluation
include, but are not limited to, subject pool access, diversity, speed, and cost (Mason
Additionally, non-response error has been found to be less of a concern than that
obtained through other Internet convenience samples (Paolacci, Chandler, & Ipeirotis,
2010). Related to cost, one researcher noted that while they “pay $8 for a 15- to 20-
minute experiment in a lab…we can run the same study on MTurk for 75 cents to a
dollar” (Bohannon, 2011, p. 307). Azzam and Jacobson (2013), for instance, paid $.75
per completed survey, and had 500 responses within three days at a total cost
(including commission) of $412.50. Because of these advantages, researchers have
used Mturk for a wide variety of studies, such as developing a tool to track the
emotional content of Twitter messages (Bohannon, 2011, p. 307), deciphering ‘bad’
handwriting in a process of iterative refinement (Little, Chilton, Goldman, & Miller,
2009), developing a screening process to identify conscientious Mturk workers
(Downs, Holbrook, Sheng, & Cranor, 2010), serving as a matched comparison group
for an evaluation (Azzam & Jacobson, 2013), and participating in a wide range of surveys (Paolacci, Chandler, & Ipeirotis, 2010; Ipeirotis, 2010).

The quality of data using Mturk has been investigated in a number of studies, with a growing number indicating that results are comparable or even superior to established alternatives (Mason & Suri, 2012, p. 4). While any effort to collect data through the Internet has clear limitations, results utilizing Mturk have been shown to be consistent with traditional approaches and provide better diversity as compared with traditional samples (Gosling, Vazire, Srivastava, & John, 2004). Test-retest reliabilities were found to compare favorably with correlations of traditional methods, and the data provided by MTurk “met or exceeded the psychometric standards associated with published research” (Buhrmester, Kwang, & Gosling, 2011, p. 5). Similarly, Mturk workers annotating affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation were found to be in high agreement with experts (Snow et al., 2008). One researcher summarized this growing body of research by saying, “…there are numerous studies that show correspondence between the behavior of workers on Mechanical Turk and behavior offline or in other online contexts…evidence that Mechanical Turk is a valid means of collecting data is consistent and continues to accumulate” (Mason & Suri, 2012, p. 4).
The Study

This study was built on the foundation of work by Azzam and Jacobson (2013) related to the potential value of utilizing the Mturk platform for various aspects of evaluation projects, and explores the potential to “use crowdsourcing to rate or even help code transcribed interview or focus group passages, audio recordings, or video recordings” (Azzam & Jacobson, 2013, p. 11). Additionally, Azzam later noted the opportunity to explore the use of Mturk for qualitative analysis (Azzam, 2013). Evaluators unfamiliar with Mturk should consult those sources for additional background and a useful overview of how crowd-sourcing could benefit evaluations.

The broad problem addressed by this study is the underutilization of visual data within the evaluation community, and the specific problem addressed is nested within that larger whole: the barrier to use associated with the difficulty of analyzing visual data within typical evaluation budgets and schedules. Crowd-sourcing is suggested as an intriguing option to address this problem. The goal of this study is to provide an initial exploration of the approach and to compare the strengths and weaknesses associated with varied coding tasks and sizes of coder cohorts. While there are multiple crowd-sourcing platforms available, the Mturk platform was utilized for the current study because it is “the largest, most well-known, and most empirically studied of these sites” (Azzam & Jacobson, 2013, p. 2-3). Three interrelated questions inform this study:

1. What is the cost, speed and agreement of crowd-sourced categorization coding of video data?
2. How do different types of categorization questions and different sizes of coder cohorts impact the results?

3. Are coding judgments biased by how personally relevant the coders consider the video data?

Method

The method for this study was developed to address the three research questions and is presented in the five categories of design, instrumentation, sample, procedure, and analytic approach.

Design

The design of the current study was developed to complement and sequentially follow a companion study (Wilson & Azzam, manuscript in preparation), which explored the use of crowd-sourced participants in the initial ‘open coding’ stage of analyzing video data. While both studies utilize Mturk and the same video data, they address distinctly different questions and stages of the coding process. The research questions of the current study led to a comparative, exploratory design that utilized the Mturk platform and compared three different sizes of coding cohorts (3, 10 and 20) and ten different coding tasks (described in the ‘instrumentation’ section). Web-based surveys were developed on the SurveyMonkey platform to operationalize the coding tasks and facilitate the engagement of unsupervised, geographically dispersed participants hired through the Mturk platform.
**Instrumentation**

In order to keep each survey-based coding task within a length typical for Mturk, a total of 10 categorization prompts and 7 demographic questions were operationalized into web-based survey questions on the SurveyMonkey platform. The survey questions were developed to explore a diversity of coding decisions and degrees of interpretive difficulty, and the specific wording and category descriptions were based on the findings from the previous open-coding study (Wilson & Azzam, Manuscript in Preparation). The prompts for the code decisions and associated relevance to the research questions for this study are as follows:

1. **Was the product reviewed something you personally might use?** This indicator of the personal relevance of the video data was a theme found in the open-coding study (Wilson & Azzam, Manuscript in Preparation). Simple ‘yes/no’ categories were used to prompt personal reflection and provide data for testing coder bias related to personal relevance.

2. **What was the gender of the presenter?** This prompt represents the variety of categorization coding decisions related to accurate identification of objectively straightforward visual data needed for data description and basic analysis.

3. **What identifying information about the product was provided?** Three categories (model only, manufacturing company only, model and manufacturing company) requiring careful attention to the entire video to identify a specific type of content. Additionally, this requires the conceptual
understanding of the categories, requiring more judgment and interpretation than the gender prompt.

4. **What features of the product were described?** Three categories (strengths only, weaknesses only, strengths and weaknesses) conceptually similar to prompt number three to facilitate comparison.

5. **What product features were mentioned?** This first of two ‘select all that apply’ coding question included three code options of product features, which were all clearly mentioned in the video data.

6. **What product features were mentioned?** This second ‘select all that apply’ coding question included at least one product feature, which was clearly mentioned and one or two (depending on video) product feature which was not mentioned to facilitate comparison with the list in prompt number five, where all were referenced.

7. **What was the quality of the video picture?** A two category question (blurry or clear) requiring subjective judgment of the visual quality. The survey was intentionally designed without any reference image for comparison to assess the consistency of unaided judgment as an indicator for the most difficult case.

8. **Did the reviewer verbally recommend the product?** Three categories, two of which include brief text as reference: no verbal recommendation, mixed verbal recommendation (i.e., “it’s o.k.” or similar), strong verbal recommendation (i.e., “I definitely recommend” or similar). This question included the qualitative short response “what words were used to recommend the product?”
to prompt careful listening and specific qualitative feedback to confirm attentiveness.

9. What emotional impact did the review have on you? Five categories for subjective personal response (i.e., strongly negative, somewhat negative, neutral, somewhat positive, and strongly positive) to correlate with personal relevance from question one, overall responses, and findings from the open-coding study (Wilson & Azzam, Manuscript in Preparation).

10. How trustworthy was the reviewer? Three subjective categories (i.e., not trustworthy, somewhat trustworthy, and very trustworthy) to correlate with personal relevance from question one, overall responses, and findings from the open-coding study (Wilson & Azzam, Manuscript in Preparation).

Sample

The sample for this study included two domains: the user-generated video data utilized as the basis for the coding decisions, and the coder participants hired through the Mturk platform. The video data coded for this study consisted of five user-generated product review videos hosted on the YouTube channel of ExpoTV.com, which is “an online community of consumers who share their unbiased, honest opinions in video” (ExpoTV, 2013). These videos were intentionally sampled to maximize the diversity of product types, demographics of the presenters, lengths, and the popularity of the videos as indicated by the number of views it had received. An overview of the five videos is provided in Table 1 below:
Table 9  
*User-Generated Product Review Videos*\(^2\) *Coded by Study Participants*

<table>
<thead>
<tr>
<th>Video #</th>
<th>Product</th>
<th>Length</th>
<th>Ethnicity</th>
<th>Gender</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cell Phone</td>
<td>1:08</td>
<td>White</td>
<td>Female</td>
<td>28,290</td>
</tr>
<tr>
<td>2</td>
<td>iPod Touch</td>
<td>2:32</td>
<td>White</td>
<td>Male</td>
<td>24,501</td>
</tr>
<tr>
<td>3</td>
<td>Board Game</td>
<td>1:42</td>
<td>White</td>
<td>Female</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Facial Moisturizer</td>
<td>3:03</td>
<td>Af. Am.</td>
<td>Female</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>Toy Helicopter</td>
<td>2:30</td>
<td>Asian</td>
<td>Male</td>
<td>2</td>
</tr>
</tbody>
</table>

While each study participant could only complete each survey once, some participants completed multiple surveys. An analysis of IP addresses of the participants indicated that 66 individuals completed 1 survey, 20 completed 2, 9 completed 3, and 4 completed 4 and 5 respectively, for a total of 103 unique participants. A similar discrepancy appeared related to the total number of surveys completed for this study; surveys 1 and 4 each had one extra completed and survey 2 had two extra completed, for a total of 169, not the 165 paid through Mturk. The likely reason for the discrepancy is that surveys could be completed on SurveyMonkey without participants taking the final step of entering the unique ‘completion code’ on the Mturk platform; participants that forget that step would register as a completed survey in SurveyMonkey, but not be paid nor counted on Mturk, resulting in additional

\(^2\) Citations for the video data are marked with an ‘*’ in the References
individual(s) being linked to the survey. The data from all 169 surveys was utilized for this study as all were equally valid and there was no way to determine which surveys were completed by workers which were ultimately unpaid.

The demographic characteristics of the Mturk participants of this study were broadly consistent with the findings from other studies (i.e. Ipeirotis, 2010) as the participants had a higher educational level than the general population and are “at least as representative of the U.S. population as traditional subject pools, with gender, race, age and education of Internet samples all matching the population more closely than college undergraduate samples and internet samples in general” (Paolacci, Chandler, & Ipeirotis, 2010, p. 414). The reliability of this self-reported demographic data is supported by another study which found the internal consistency of self-reported demographics on Mechanical Turk very high, noting only 1 worker out of 207 (0.4%) changed an answer on gender, age, education or income (Mason & Suri, 2012). In order to maintain consistency with the findings and facilitate comparison between cohorts, the demographics presented in Table 2 below are based on the 169 total surveys.

Table 10

<table>
<thead>
<tr>
<th>Participant Demographics as Percentage of Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Category</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Female</td>
</tr>
</tbody>
</table>

76
Table 10 - continued

<table>
<thead>
<tr>
<th>Demographic Category</th>
<th>3 Coders</th>
<th>10 Coders</th>
<th>20 Coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>33%</td>
<td>76%</td>
<td>54%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>47%</td>
<td>68%</td>
<td>36%</td>
</tr>
<tr>
<td>30-49</td>
<td>33%</td>
<td>20%</td>
<td>37%</td>
</tr>
<tr>
<td>50 and older</td>
<td>20%</td>
<td>12%</td>
<td>25%</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school graduate</td>
<td>13%</td>
<td>14%</td>
<td>15%</td>
</tr>
<tr>
<td>Trade/vocational training</td>
<td>0%</td>
<td>12%</td>
<td>6%</td>
</tr>
<tr>
<td>Some college</td>
<td>27%</td>
<td>34%</td>
<td>33%</td>
</tr>
<tr>
<td>College graduate</td>
<td>40%</td>
<td>30%</td>
<td>31%</td>
</tr>
<tr>
<td>Some postgraduate work</td>
<td>0%</td>
<td>2%</td>
<td>4%</td>
</tr>
<tr>
<td>Post graduate degree</td>
<td>20%</td>
<td>8%</td>
<td>10%</td>
</tr>
<tr>
<td>Employment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed full time</td>
<td>40%</td>
<td>34%</td>
<td>38%</td>
</tr>
<tr>
<td>Employed part time</td>
<td>40%</td>
<td>16%</td>
<td>32%</td>
</tr>
<tr>
<td>Unemployed</td>
<td>13%</td>
<td>46%</td>
<td>24%</td>
</tr>
<tr>
<td>Retired / Disabled</td>
<td>7%</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>Marital</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 10 - continued

<table>
<thead>
<tr>
<th>Demographic Category</th>
<th>3 Coders</th>
<th>10 Coders</th>
<th>20 Coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single/never been married</td>
<td>47%</td>
<td>72%</td>
<td>59%</td>
</tr>
<tr>
<td>Married</td>
<td>53%</td>
<td>16%</td>
<td>19%</td>
</tr>
<tr>
<td>Divorced</td>
<td>0%</td>
<td>8%</td>
<td>14%</td>
</tr>
<tr>
<td>Separated / widowed</td>
<td>0%</td>
<td>2%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Race

<table>
<thead>
<tr>
<th>Race</th>
<th>3 Coders</th>
<th>10 Coders</th>
<th>20 Coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>87%</td>
<td>72%</td>
<td>61%</td>
</tr>
<tr>
<td>Black or African-American</td>
<td>0%</td>
<td>4%</td>
<td>20%</td>
</tr>
<tr>
<td>Asian</td>
<td>13%</td>
<td>16%</td>
<td>4%</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>0%</td>
<td>2%</td>
<td>7%</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td>Other</td>
<td>0%</td>
<td>6%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Procedure

This study was conducted following a companion study, which explored crowd-sourcing the descriptive or ‘open-coding’ aspects of video analysis (Wilson & Azzam, Manuscript in Preparation), and both the experience of conducting the previous study and findings generated informed the development of the current study in important ways. While Mturk provides an HTML based interface with a selection of templates for common HITs, as well as freedom for extensive customization through a Java based ‘application programming interface,’ neither were used for this project.
because of the associated additional technical complications. Instead of developing the survey directly on the Mturk platform, one web based survey per video (for a total of 5) was developed on SurveyMonkey.com. These surveys included the product review video data as an embedded file to maximize ease of use. Links to these surveys were embedded within an existing Mturk template for distribution to Mturk participants.

Following a pilot test and minor text revisions for clarification, the surveys were distributed through Mturk to three cohorts of 3, 10, and 20 workers in sequence. To increase the anticipated diversity of participants, the surveys were distributed at different times of the day over the course of five days. The Mturk workers were paid between $.50 and $.75 depending on the length of the video, and were qualified based on criteria requiring that they be located in the United States and have completed at least 500 approved HITs with an approval rating of 90%.

**Analytic Approach**

The data for this study consisted of the coding decisions completed by the Mturk participants. This coding (organized as survey responses) was downloaded from SurveyMonkey and analyzed in MS Excel. Six of the coding questions involved accurately identifying video content that was objectively present or absent, including the gender of a given video’s presenter, product identification, feature overview, Product List 1 and 2, and verbal endorsement. The ‘correct’ coding decision for each of these was determined by the principal researcher based on multiple viewings of the videos. One of the coding questions involved a subjective determination of video
quality (blurry or clear) and the ‘correct’ coding decision for this was based on the findings from the crowd-sourced open coding study utilizing the same videos (Wilson & Azzam, manuscript in preparation). The analysis conducted for these coding decisions involved determining the level of agreement with the correct categorization, calculated as a percentage agreement, similar to grading a multiple choice test. These percentages were first calculated at the cohort level to facilitate comparison of agreement between both different types of coding decision and different size cohorts. The findings of this analysis are presented in Tables 3, 4, and 5. The percentages of agreement were also calculated at the video level, a procedure, which involved combining the agreement from the seven question types for each of the five videos. This analysis facilitated the comparison of agreement between different videos. The findings of this analysis are presented in Table 6.

The analytic approach for the third research question (i.e., coding judgments being biased by how personally relevant the coder considers the video data) was to determine if there was any correlation between the participant likelihood of using the product featured in the video data and the two subjective coding decisions (i.e., emotional impact of the video, trustworthiness of the presenter). This correlation was determined by calculating two Pearson’s Product-Moment Correlation Coefficients (r): one between the likelihood of using the product and the emotional impact of the video, and the second between likelihood of using the product and the trustworthiness of the presenter. The findings from this analysis are presented in Table 7.
Findings

The findings are presented in three sections corresponding with three types of comparative analysis and roughly corresponding with the research questions. As it didn’t involve comparative analysis, the cost and speed of conducting this study as relates to the first research question is addressed in the discussion, not findings. The agreement of crowd-sourced categorization coding and different types of categorization questions and different sizes of coder cohorts related to questions one and two are addressed in the findings through two levels of analysis. First, the cohort analysis presents the findings of the agreement calculated at the cohort level. Second, the video analysis presents the findings calculated at the video level. The final analysis presented addresses the third question related to the potential bias in coder judgment, and the analysis of bias section presents the findings of a correlation analysis.

Cohort Analysis

The comparison of average agreement by cohort presented in table 3 reveals that the only coding question showing perfect consistency was the question about the gender of the presenter in the 3 coder cohort, followed by the gender question in the 10 and 20 coder cohorts each with 98% agreement (see column one of Table 3). This is intuitive, as it involves a very simple, binary coding decision, and the smaller number of coders reduces the risk of mistakes or inattentive coders. Overall Product List 1 (the list of product features which were all present) was the second highest, with all
three cohorts above 90%, and well above the agreement for Product List 2 (the list of product features where some were missing) which ranged from 73% to 85%. The similar questions related to the ID info, Feature Overview and Product Endorsement all showed a similar range of 62% to 80%, with no clear pattern between the cohorts.

Table 11

Overview of Average Agreement by Cohort

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Gender</th>
<th>ID</th>
<th>Feature Info</th>
<th>Feature Overview</th>
<th>Product List 1</th>
<th>Product List 2</th>
<th>Video Quality</th>
<th>Verbal Endorse</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100%</td>
<td>80%</td>
<td>87%</td>
<td>91%</td>
<td>73%</td>
<td>73%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>98%</td>
<td>74%</td>
<td>62%</td>
<td>93%</td>
<td>85%</td>
<td>64%</td>
<td>78%</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>98%</td>
<td>78%</td>
<td>71%</td>
<td>94%</td>
<td>77%</td>
<td>52%</td>
<td>78%</td>
<td></td>
</tr>
</tbody>
</table>

The next highest rate of agreement for the three coder cohorts (with an average of 91%) was Product List 1, requiring the identification of specific product features, which were all present. The percentage correct for identifying product features when one or more was ‘not present’ in the list (Product List 2) is a more rigorous test of coding agreement and dropped to 73% for the 3 coder cohort, equal to the agreement for the question involving the degree of verbal endorsement. The product ID and overall feature list coding questions (80% and 87%, respectively) were relatively high and involve similar three code decisions. The subjective video quality code was comparable to others, and similar to other cohorts, was perfect for the toy helicopter video.
As shown in the first column of Table 5 regarding findings for the 10 coder cohort, one coder incorrectly coded the gender of the presenter for the board game video, but otherwise there was perfect agreement on that simplest coding question. The feature overview (strengths, weaknesses, or both) presented the overall lowest agreement for this cohort with 62%, impacted especially by the radio controlled toy and board game videos, and over 10% lower than the similar ID info. Like the three coder cohort, Product List 1, where all features were present, was the second highest at 93%, and Product List 2, involving omitted features, was clearly lower at 85%. While the video quality coding showed generally low agreement, there was 100% agreement
on the coding of the radio controlled toy, consistent with the fact that it appeared to be the only video filmed on a high definition camera and had the sharpest image quality of the five videos overall.

Table 13

<table>
<thead>
<tr>
<th>Video</th>
<th>Gender</th>
<th>ID</th>
<th>Feature Info</th>
<th>Feature Overview</th>
<th>Product List 1</th>
<th>Product List 2</th>
<th>Video Quality</th>
<th>Verbal Endorse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Phone</td>
<td>100%</td>
<td>90%</td>
<td>80%</td>
<td>83%</td>
<td>90%</td>
<td>30%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>iPod Touch</td>
<td>100%</td>
<td>90%</td>
<td>70%</td>
<td>97%</td>
<td>80%</td>
<td>40%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Board Game</td>
<td>90%</td>
<td>80%</td>
<td>20%</td>
<td>93%</td>
<td>83%</td>
<td>70%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Moisturizer</td>
<td>100%</td>
<td>90%</td>
<td>100%</td>
<td>97%</td>
<td>87%</td>
<td>80%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>Toy Helicopter</td>
<td>100%</td>
<td>20%</td>
<td>40%</td>
<td>97%</td>
<td>87%</td>
<td>100%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>98%</td>
<td>74%</td>
<td>62%</td>
<td>93%</td>
<td>85%</td>
<td>64%</td>
<td>78%</td>
<td></td>
</tr>
</tbody>
</table>

As shown in the first column of Table 6 about the 20 coder cohort, one respondent incorrectly coded the gender of the presenter for both the moisturizer and radio controlled toy videos, but otherwise the cohort of 20 coders showed perfect agreement on the simplest of coding decisions. The overall average rate of agreement (52%) for the subjective code of video quality was the lowest for any cohort, but similar to the 3 and 10 coder cohorts, as there was nearly perfect agreement (only one coder of 20 wrong) for the toy helicopter video. Similarly, Product List 1, where all
features were present, was the second highest at 94%, and the Product List 2, involving omitted features, was markedly lower at 77%. The remaining three codes (ID Info, Feature Overview, and Verbal Endorsement) each involved three categories, and showed similar agreement in the range of 70%.

Table 14

20 Coder Cohort Agreement

<table>
<thead>
<tr>
<th>Video</th>
<th>Gender</th>
<th>ID Info</th>
<th>Feature Overview</th>
<th>Product List 1</th>
<th>Product List 2</th>
<th>Video Quality</th>
<th>Verbal Endorse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Phone</td>
<td>100%</td>
<td>100%</td>
<td>76%</td>
<td>95%</td>
<td>73%</td>
<td>29%</td>
<td>24%</td>
</tr>
<tr>
<td>iPod Touch</td>
<td>100%</td>
<td>91%</td>
<td>68%</td>
<td>98%</td>
<td>70%</td>
<td>36%</td>
<td>100%</td>
</tr>
<tr>
<td>Board Game</td>
<td>100%</td>
<td>90%</td>
<td>45%</td>
<td>92%</td>
<td>80%</td>
<td>35%</td>
<td>95%</td>
</tr>
<tr>
<td>Moisturizer</td>
<td>95%</td>
<td>86%</td>
<td>90%</td>
<td>89%</td>
<td>84%</td>
<td>67%</td>
<td>95%</td>
</tr>
<tr>
<td>Toy Helicopter</td>
<td>95%</td>
<td>25%</td>
<td>75%</td>
<td>93%</td>
<td>80%</td>
<td>95%</td>
<td>75%</td>
</tr>
<tr>
<td>Average</td>
<td>98%</td>
<td>78%</td>
<td>71%</td>
<td>94%</td>
<td>77%</td>
<td>52%</td>
<td>78%</td>
</tr>
</tbody>
</table>

Video Analysis

As shown in Table 7, the variation in coding agreement was relatively small between the different videos, with only a 19% difference between the highest and lowest. Despite the broad similarity, an underlying pattern is apparent, as the cell phone coding had the lowest rate of agreement in each of the three cohorts, and the moisturizer was the highest in both the 10 and 20 coder cohorts and had the highest
overall average. A more specific analysis of the underlying data related to these videos revealed that these differences are related to coding agreement on just two coding decisions: video quality and verbal endorsement. In each case, the rate of agreement for these two was unusually low (around 30%) for the cell phone and average or above average for the moisturizer (ranging from 67% to 90%), and this difference accounted for the pattern. For the cell phone video, the quality was considered blurry and the presenter did not explicitly endorse the product, and for the moisturizer, the video quality was clear and the presenter did verbally endorse the product, indicating that the crowd-sourced workers tended to inaccurately identify negative cases, similar to the inaccuracies for Product List 2.

Table 15

Average Agreement by Video and Coder Cohort

<table>
<thead>
<tr>
<th>Video</th>
<th>3 Coders</th>
<th>10 Coders</th>
<th>20 Coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Phone</td>
<td>79%</td>
<td>72%</td>
<td>71%</td>
</tr>
<tr>
<td>iPod Touch</td>
<td>86%</td>
<td>82%</td>
<td>81%</td>
</tr>
<tr>
<td>Board Game</td>
<td>81%</td>
<td>77%</td>
<td>77%</td>
</tr>
<tr>
<td>Moisturizer</td>
<td>78%</td>
<td>90%</td>
<td>87%</td>
</tr>
<tr>
<td>Toy Helicopter</td>
<td>89%</td>
<td>75%</td>
<td>77%</td>
</tr>
</tbody>
</table>
**Analysis of Bias**

Table 8 presents the Pearson’s r coefficients of correlation for all participants (i.e., combining the 3, 10 and 20 cohorts) between personal relevance and the coding of the trustworthiness of the presenter and the overall emotional impact of the video on the viewer. As the overall coefficient of .05 indicates, there was very little correlation related to coding the trustworthiness of the presenter. The higher, but nonetheless still small, correlation of .23 with the emotional impact of the video on the viewer does suggest a small relationship between personal relevance and emotional impact, but the overall low levels of correlation indicate that the personal relevance of the data reviewed did not bias the subjective and emotive coding decisions.

**Table 16**  
*Correlation Between Personal Relevance and Emotive Coding*

<table>
<thead>
<tr>
<th>Video</th>
<th>Trustworthy Presenter</th>
<th>Emotional Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Phone</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>iPod Touch</td>
<td>-0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>Board Game</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td>Moisturizer</td>
<td>0.16</td>
<td>0.40</td>
</tr>
<tr>
<td>Toy Helicopter</td>
<td>-0.04</td>
<td>0.17</td>
</tr>
<tr>
<td>Overall</td>
<td>0.05</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Note: *Pearson’s Product-Moment Correlation Coefficient (r).*
In addition to the correlation coefficient analysis, qualitative analysis of the underlying data of the open-coding study revealed that the findings of the present study related to personal relevance, trustworthiness of the presenter, and overall emotional impact were generally consistent with the qualitative data from the companion study (Wilson & Azzam, manuscript in preparation). In specific, this analysis revealed that the videos most frequently coded in the present study as having both the most emotionally impactful and most trustworthy presenter (facial moisturizer) and least emotionally impactful and least trustworthy presenter (cell phone) were both reflected in the data provided in the open-coding study (Wilson & Azzam, manuscript in preparation). Additional research is needed to confirm these findings, but the similarities in data between two studies and two groups of participants indicates that subjective coding decisions can be stable within larger cohorts.

**Discussion**

The findings of this study indicate that crowd-sourced workers can provide very reliable coding within specific parameters, and that further refinements to the methodology are needed to demonstrate reliable results for a more diverse set of data and coding decisions. The coding for gender showed perfect or nearly perfect agreement across all coding cohorts, indicating that objectively clear and simple coding can be provided reliability with even very small coding cohorts. The very high rates of agreement for coding video quality of the radio controlled toy (equaling the
reliability of gender) indicates that crowd-sourced workers can also provide consistent coding even for more subjective decisions. That said, the high variability and generally low level of agreement in coding video quality for the other videos indicates that more detailed categorization formation, comparison images, or other preparation materials should be considered to achieve high agreement on a wider range of videos. The generally high agreement on product features which were present and other ‘positive cases’ (i.e. ID Info, Feature Overview, Product List 1) indicates that crowd-sourced workers can identify specific content details. At the same time, the fact that ‘negative case’ examples were consistently lower suggests that worker pre-qualification or additional instruction should be considered for reliably coding finer content details.

The unsupervised and geographically dispersed nature of crowd-sourced workers does not necessarily lead to carelessness and confusion. Quite to the contrary, this study indicates that Mturk workers can be attentive and provide useful work that is both time and cost effective, as discussed below. The quality and reliability of crowd-sourced coding, similar to traditional team projects, is largely related to the clarity of communication and care in formulating the coding categories – broader leadership responsibilities that can rarely be even delegated, much less outsourced or crowd-sourced. As such, the responsibility of the principal researcher or analyst utilizing Mturk workers shifts from direct coder to manager and designer, and attentiveness to the important details within that role is clearly critical.
Potential Benefits

One of the primary reasons video data is neglected in evaluation and research project is the reality of the time (and associated cost) of analysis (Luff & Heath, 2012), and this study demonstrated that crowd-sourcing can effectively address both. As relates to time, while there is additional work associated with learning and developing the tools on the SurveyMonkey and Mturk platforms, once those initial steps are completed, the actual coding was completed very quickly by the Mturk workers; within 15 minutes for the 3 coder cohort and within 4 hours for the 10 and 20 coder cohorts. Because of the development time, this methodology would not likely be time-effective for small projects but could be very effective for large projects where the speed gained during the coding far surpasses the extra time spent in developing the platforms. Additionally, if the online platform could be completed before collection of the primary bulk of video data (i.e., after completion of a small pilot or when collecting similar data in multiple locations) the extreme speed of crowd-sourced coding could facilitate initial descriptive reporting literally overnight. As for cost, the 165 completed surveys (or HITS) of the present study cost a total of $117.98 (Mturk participants plus Amazon commissions), or $0.55 for the shorter videos and an average of $0.715 per completed survey. While this cost could likely have been reduced even further by offering lower rates for each survey, this study sought to avoid the ‘digital sweatshop’ ethical concerns related to Mturk workers (Scholz, 2012). Despite these good intentions, the effective hourly rate ultimately paid Mturk participants for this study ranged from $5.23 to $9.28, which given the US Federal
minimum wage of $7.25 (United States Department of Labor, 2013, paragraph 1) indicates that the rates offered were arguably too low in some cases.

**Limitations and Future Research**

This study was designed to explore the viability of crowd-sourced categorical coding as a compliment to a similar study exploring crowd-source open coding, (Wilson & Azzam, manuscript in preparation) and as such must be understood in the context of several important limitations. First, the use of the user generated product review videos for both studies is such that other types of video data relevant to evaluators (i.e., interviews, classroom interactions, training sessions, etc.) were not addressed. Similarly, the present study was designed to explore a diversity of categorical coding tasks, and provided initial indicators across a broader range of situations, findings to be distinguished from strong evidence related to reliability of specific coding tasks. To address these limitations, additional research is needed to identify the contexts and procedures to generate broadly applicable, stable and accurate coding. For example, coding for gender was demonstrated to be highly reliable in all three cohorts, and future studies could test the reliability of other basic demographic or contextual factors such as race, age category, or setting. Similarly, the video with the clearest picture quality (toy helicopter) was coded with very high reliability across all cohorts, but the other videos demonstrating varying levels of camera quality, lighting and video compression were highly variable, and future research could compare the use of comparison images, worker training or screening, or
a wider range of categories related to audio or video quality. Additionally, future research addressing video quality could gather data related to internet connection speed, browser software and screen resolution to facilitate deeper analysis of perceived video quality. While the coding related to personal judgments of emotion (trustworthiness of the presenter, overall emotional impact) showed very little evidence of bias and general correspondence with the open coding study (Wilson & Azzam, manuscript in preparation), specific and rigorous tests of the reliability of emotive coding should be conducted. And finally, the coding related to details of the video content (which generally comprises the core goal of video analysis projects) was highly variable across the different videos, cohorts and coding questions. Coding for the features of the product that were present (Product List 1) was over 90% reliable within each cohort, but the reliability of coding for product features where one or more was missing (Product List 2) dropped to between 73% and 85%, and other coding questions were even more variable. Additional research should identify the number of code categories, code descriptions, training and preparation of coders, and other design features needed to consistently produce reliable coding of the details of video data content.

Applications of Method

Three categories of application are suggested as inspiration for the needed additional research noted above and as food for thought as evaluators or researchers consider their own context.
1. Initial video data sorting: The perfect reliability of the gender coding for the 3 coder cohort indicates that utilizing small groups of coders for the initial review and sorting of video data could be extremely time and cost effective. Hours of footage broken into hundreds of short clips could be quickly reviewed and tagged to provide basic descriptive information, such as the gender, race, setting (i.e., inside or outside), language, or broad age category (i.e., child, teen or adult). Similarly, the equally high reliability in categorizing the video quality for the radio controlled toy indicates that it is possible for crowd-sourcing to accurately categorize the quality of the picture in some cases, but the very high variance between the videos also indicates that comparison samples or additional instruction would be required for accurate coding in the situations where the quality is less than perfect. If combined, crowd-sourcing this range of basic descriptive codes could effectively help evaluators or researchers quickly identify video data with the desired content or production value to guide additional analysis or provide overall summaries of large datasets.

2. Analyzing existing video data: A substantial volume of video data either produced by or related to many programs, organizations or educational evaluands is already uploaded on organization websites or YouTube. This crowd-sourced coding method could be a time and cost-effective method for initial analysis of this largely untapped source of data and could provide useful insights related to either the context, public perception, or operation of the evaluand.

3. User-generated evaluation videos: While certainly not appropriate for all evaluations, in many situations program participants or appropriate stakeholders might
record videos with evaluative feedback. In addition to video data for evaluations, these videos could be a valuable source of data with several benefits not available through traditional surveys or interviews, such as teaching, promotion or organizational learning. The recordings could be conducted at home or work on webcams, during meetings or events at kiosks or facilitated by camera operators, as part of a contest, on cell phones during training events, or any number of other creative means. Once recorded, the video data could be analyzed utilizing the crowd-sourcing methods explored in both this study and the related open-coding study (Wilson & Azzam, Manuscript in preparation).
IV. RELATIONAL INTEGRITY: A CONCEPTUAL FRAME FOR ETHICAL VISUAL METHODOLOGIES INFORMED BY THE FIELDS OF ENTERTAINMENT, JOURNALISM, ADVERTISING AND SOCIAL MEDIA

Authors

Kurt Wilson, Western Michigan University, Patricia Reeves, Western Michigan University and Daniela Schroeter, Western Michigan University

Abstract

Using photographs or video as data raises ethical questions many evaluators feel unprepared to address, and this hesitation and confusion is a barrier to the more widespread use of this important source of data. This paper offers the conceptual framework of ‘relational integrity’ identified as a theme in the literature of ethical visual methodologies as a compliment to the Code of Research Ethics and Guidelines published by the International Visual Sociology Association (Papademas, 2009). The implications and potential benefits of this approach are explicated for the especially challenging ethical domains of informed consent and confidentiality/anonymity, with particular focus on the parallels between this approach and ‘contextual integrity’ as has been advanced within the legal community.

Key themes and findings from a comparative analysis of the image use procedures in the fields of entertainment, journalism, advertising, and social media are presented as contextual reference for guiding ethical decisions and informing relationships developed surrounding visual methodologies. Five categories of
relationship are suggested based on this analysis to serve as reference for evaluators. An inductive analysis of the assumptions and structures leading to the image use procedures indicated a considerable cultural divide between the more scholarly evaluation community and the broader public related to visual methodologies.

Keywords: ethics; ethical; visual methodology; photo; video; journalism; advertising; entertainment; internet; relational integrity

Introduction

The use of photographs or video as data raises ethical questions many evaluators, researchers, and IRB committees feel unprepared to address (Allen, 2009). Policy and guidelines are often either missing or conflicting (Carusi & Jirotka, 2009), and this creates fear and confusion that can become a barrier to the more widespread use of this important source of data (Allen, 2009, Lincoln & Tierney, 2004). Given the fast pace of technological change, critical issues such as legal requirements, storage, access, and security of data can be overwhelming, such that “…a case could be made by those skeptical about the visual that the incorporation of visual methods is literally more trouble than it is worth” (Prosser & Loxley, 2008 p. 50). While these issues might be relatively unfamiliar within many evaluation and research contexts, regular use of images has been commonplace in the fields of entertainment, journalism, and advertising for decades, and common image use themes are emerging for internet use as well (Smith, 2013).
This study was developed in the context of the International Visual Sociology Association Code of Research Ethics and Guidelines, a document which helps researchers and evaluators to “…be aware of ethical issues in the research process, to encourage individual responsibility for ethical practice, and to provide a supportive document for visual researchers pursuing formal approvals from ethics review boards, academic institutions, and prospective sponsors” (Papademas, 2009, p. 251 - 252).

More specifically, the goal of this study is to provide a conceptual framework and resource that compliments this code by providing “practical resources to guide us in negotiating the moral complexities of visual communication” (Perry, & Marion, 2010, p. 96).

Context for Visual Methods

History and Culture

Visual methods have a long history; photography was invented in 1798 (Gernsheim, 1986) and Thomas Edison patented the first movie camera in 1888 (Green, 2013). The methods for capturing still and moving images have been progressively refined and simplified ever since. The trend toward broader use of photography especially accelerated in recent years with low-cost digital cameras on cellphones; by 2010, over one billion camera phones were in use worldwide (The Economist, 2010). As one indicator of the current volume of images recorded and viewed, 300 million photos are uploaded to Facebook every day, the total archive of photos uploaded to Facebook exceeds 240 billion photos, and these are viewed at a
rate of several million per second (Madrigal, 2013). These numbers are especially significant when considering that Facebook was launched as a student project less than 10 years ago (Tabak, 2004), indicating both rapid technologic adoption and considerable latent demand for visual communications. Similarly, YouTube was launched in just 2007, and currently has over 1 billion unique visitors watching 6 billion hours of video each month and uploading new video content at a rate of over 100 hours of footage every minute (YouTube, 2013). These references underscore two points relevant for the context of this study: recording still and moving images has a long history, and visual communications are a substantial part of the life and cultural context for a very large number of people.

While unfamiliarity might lead some to consider visual research methods as ‘new,’ the historic record indicates otherwise. Photography and film have been used in sociology and anthropology since the 1920’s (Shrum, Duque & Brown, 2005), and the community surrounding the use of visual methods has organized both the International Visual Sociology Association (IVSA) and Commission on Visual Anthropology (CVA): The IVSA hosts annual conferences and publishes a peer-reviewed journal called Visual Studies (IVSA, 2013), and the CVA hosts its own program within the conferences of the International Union of Anthropological and Ethnological Sciences and publishes Visual Anthropology (CVA, 2010). While specific technologies have changed substantially, “…the fundamental considerations and theoretical choices involved in conducting qualitative research have not changed significantly since the 1960s…after learning the new technology, one still has to engage with familiar issues
and problems” (Travers, 2009, p. 171-172). While the broad domain of ethics is one category of ‘familiar issues and problems’ that researchers using visual methods must engage, the considerations and questions within that domain are often substantially different than those raised by numeric or textual data (Wiles et al., 2008, p. 8). The ethical issues raised by images often seem uniquely fraught because, “whilst texts are associated with reason and higher mental faculties, images are seen as subversive, dangerous and visceral” (Emmison & Smith, 2000, p. 14). Additionally, the ethical issues surrounding visual methods are often seen as substantially more complex, such that “…the common pessimistic presumption that graphic media are so slippery and completely heterogeneous as to elude any kind of ethical governance or best practice” (Perry & Marion, 2010, p. 98).

**Primary Domains of Ethical Visual Methodology**

The domains of informed consent and privacy (and related issues of confidentiality and anonymity) were particularly noted in the literature as raising unique challenges for the ethical use of visual methods (Bennington, Gay, & Jones, 1999; Heath, Hindmarsh, & Luff, 2010; Allen, 2009). For many ‘traditional’ evaluation or research projects, the process of securing informed consent is fairly straightforward, generally requiring research participants to read and sign a form approved by an Institutional Review Board (IRB) or similar entity. The language on these forms balance ethical, legal, and practical concerns often developed in a forge of heated controversy, such that projects introducing variables and raising concerns that
fall far outside the typical scope can be unwelcome. A particular challenge of establishing informed consent for visual methodologies is clearly defining what consent is given for, as important options related to visual data must be considered at each step of the process: data collection, analysis, archiving/future use, and reporting. For example, gathering existing still photos (e.g., from a family album) involves different ethical questions than when the evaluator is the photographer, which is different still from participant photography. The added dimensions of audio and action add yet more important considerations for the use of video data. The extent and geographic distribution of the analysis team, the duration the materials will be archived and if they will be made available to other researchers, and the nature and context of how they will be used in reporting also must be considered, each with potential for nuances that are unique to visual methods (Wiles et al, 2008). Because of these complexities, “…there is no consensus amongst researchers concerning how to know when enough has been done to achieve informed consent, and when the point of doing too much has been reached” (Wiles, Crow, Charles & Heath, 2007, p. 5). To make matters worse, “several researchers noted that however extensive the information provided about a study is, there is still scope for misunderstanding” (Wiles, Crow, Charles & Heath, 2007, p. 5).

While various definitions for privacy, confidentiality and anonymity abound, for the purposes of this study, “privacy” is the control the participant has over the extent, timing, and circumstances of sharing information about themselves (American Journal of Evaluation, 2005, p. 357) such that “…private information about
individuals collected is understood to be private information when an individual can reasonably expect that the information will not be made public with personal identifiers” (Papademas, 2009, p. 254).

Just as visual methodologies are often misunderstood as ‘new,’ so too is the notion that concern for privacy developed only in our current age of electronic communications and government surveillance. Again, the historic record indicates otherwise, as privacy concerns related to communication practices have been discussed and debated for well over a century. In their landmark paper on privacy published in 1890 in the Harvard Law Review, Warren and Brandeis stated that “…modern enterprise and invention have, through invasions upon his privacy, subjected him to mental pain and distress, far greater than could be inflicted by mere bodily injury….to determine in advance of experience the exact line at which the dignity and convenience of the individual must yield to the demands of the public welfare or of private justice would be a difficult task…” (Warren & Brandeis, 1890, p. 8).

Confidentiality relates to the information disclosed in the context of the relationship (American Journal of Evaluation, 2005, p. 357) between the evaluator or researcher and participant, and anonymity is based on the concept of being “nameless or lacking individuality” (Merriam Webster, 2013).

Visual research methods present inherent challenges in each of these areas, and many of the common approaches used within the social sciences to provide anonymity and confidentiality (largely developed for textual data) are deeply problematic (Wiles
et al, 2008). In text based research, anonymity is established by using pseudonyms, but this approach is obviously impossible for most visual research (Prosser & Loxley, 2008 p. 54). While some might assume that the technique of pixelating the face or eyes is a suitable equivalent, in many situations this is a cure worse than the disease; such measures are rarely sufficient to ensure anonymity (because of remaining cues of context/voice), may defeat the purpose of the photo or video in the first place because of lost data related to facial expressions, and discriminates against respondents who have freely chosen to be seen (Prosser & Loxley, 2008). Even more concerning is the way that blurring or obscuring faces “objectifies people and removes their identity…without faces people appear not as people at all but as objects, this does not accord with a duty to treat people with respect…additionally pixilation of images has associations with crime; it is a commonly used device in the media when talking with ‘criminals’ or ‘victims’ of crime” (Wiles et al., 2008, p. 24). In addition to these difficulties related to the data collection, as relates to the analysis, within the dissemination and archiving phases of the process, “the ease with which images of people are copied, transformed and posted on through cyberspace ensures that no guarantee of anonymity and confidentiality can be made” (Prosser & Loxley, 2008 p. 53).

**Relational Integrity**

While the importance of the relational dimension of research is obviously not unique to visual methods (e.g., Wiles et al, 2008, Banks, 2001), it seems especially
promising as a unifying framework for addressing the diversity of ethical questions related to visual methods noted above. “Explicitly valuing participants and recognizing the potential interpersonal impact of the inquiry helps demonstrate that the researcher will be deeply ethical” (Marshall & Rossman, 2011, p. 50). Toward that end, while some research projects require laboratory conditions and professional distance implicit in the term ‘research subjects,’ the more personal and evocative engagement inherent with most visual methodologies requires a more relational collaboration with ‘research participants’ (Wiles, Crow, Charles & Heath, 2007). Similarly, several researchers noted how they prioritized the relationship with participants by allowing them to retain control of data in which they are represented (e.g., Carusi & Jirotka, 2009, p. 293, Johnson, 2011). Another researcher noted, “…even in cases where permission is granted to undertake recording, we provide the participants with successive opportunities to discuss their reservations or concerns. This helps to develop trust and a working relationship with the participants” (Heath, Hindmarsh, & Luff, 2010, p. 17, emphasis added).

Within a relational approach, the goal of informed consent is achieved not “merely as some procedural hoops that one must jump through” (Marshall & Rossman, 2011, p. 44), but rather is established, nurtured and maintained through an ongoing relational process (Fang & Ellwein, 1990). “The key to the consent process is…the development of trust by those who are video recorded that the investigators will not abuse the privilege of being allowed to acquire video data for research purposes” (Mackenzie, 2004, p. 196). Similarly, the relevant details of the informed
consent process should be explained in accessible language tailored for all participants, as “…doing research with children is not in theory different than doing it with adults because it’s all about respect whoever it is…. there’s no point in talking to adults in language they don’t understand either” (Wiles, Crow, Charles & Heath, 2007). A potentially beneficial alternative for some video projects to consider is “rather than written consent, filmed or audio-recorded consent may be more appropriate” (Wiles et al., 2008, p. 15). For example, video recording the consent would retain the personal, discussion-based context and not introduce the more legal or technical medium of a consent form, which might be especially helpful in contexts with young or illiterate participants.

While some visual methodologies can operate within the parameters of anonymity and confidentiality established for text or numeric based projects, the conceptual framework of relational integrity offers a different approach toward ethical behavior for the majority which could not. For example, the decisions of marginalized populations who have argued for the right to be made visible can be honored (Wiles et al, 2008, Carusi & Jirotka, 2009, Allen, 2012) instead of just blind insistence on the principle of anonymity. Indeed, there are many other people beyond marginalized populations that seek to be photographed, as “people in industrialized countries seek to have their photographs taken—feel that they…are made real by photographs” (Sontag, 1973, p. 161). Clearly, insuring the fully informed consent of participants and/or legal guardians is critical and could involve various complications. However, if consent is appropriately granted and the publication of images is the express will of the
participants, insisting on confidentiality and anonymity could in fact be an unethical, a paternalistic violation of the participants’ determination of how the visual data they contributed is used and disseminated. As with most research decisions, implementing the principles of anonymity and confidentiality must be appropriate to the context, if at all. As one researcher noted, in some contexts the principles of confidentiality and anonymity might not be relevant, such as “observations in public places, activities conducted in public, or other settings where no rules of privacy are provided by law or custom. Similarly, confidentiality is not required in the case of information available from public records” (Papademas, 2009, p. 254). The flexible, yet, contextually demanding yardstick of maintaining the integrity of the relationship with participants is ultimately a superior guide for ethical visual practice than a rigid insistence on the principles of confidentiality and anonymity.

The need to maintain relational integrity is also a theme in the literature related to the analysis and archiving phases of research projects. For example, one researcher referenced that during the transfer of “visual information to other persons or organizations, researchers obtain assurances that the recipients of the records, data, or information will employ measures to protect confidentiality at least equal to those originally pledged” (Papademas, 2009, p. 255, emphasis added). Other researchers working on a project involving the ethical minefield of video-taping a family with children in their home ultimately addressed the myriad ethical questions by essentially working to maintain relational integrity. They stated, “through our encouragements, we made the participants their own gatekeepers of what they felt were too private to be
filmed or observed” (Aarsand & Forsberg, 2010, p. 261). While this does not provide specific answers for every situation, the conceptual frame of relational integrity is consistent with other ethical approaches which allow “social researchers to adopt a ‘situational relativist’ approach in which ethical decisions are made on the basis of issues applicable to individual research projects” (Wiles, Crow, Charles & Heath, 2007, p. 9).

This theme of relational integrity traced through the literature above is closely related (and was inspired by) an article outlining the concept of ‘contextual integrity’ in the Washington Law Review (Nissenbaum, 2004). While the specific concern of that article was the legal tensions between privacy and government surveillance, many aspects of the argument relate to ethical tensions within evaluation and research projects. For example, researcher concerns related to data analysis, archiving and dissemination echo the concern with the “prospect of local access giving way to global broadcast” (Nissenbaum, 2004, p. 2). Because of these parallels, just as the Nissenbaum article’s “central contention is that contextual integrity is the appropriate benchmark of privacy” (p. 2) so too the contention of this paper is that relational integrity is the appropriate benchmark for visual research ethics. In an interesting example of interdisciplinary cross-pollination, it should be noted that the concept of contextual integrity was not originally a legal concept, but instead developed by social theorists seeking a more nuanced conceptual framework for privacy than the simplistic and dichotomous approach of just public and private:
“Observing the texture of people's lives, we find them not only crossing dichotomies, but moving about, into, and out of a plurality of distinct realms. They are at home with families, they go to work, they seek medical care, visit friends, consult with psychiatrists, talk with lawyers, go to the bank, attend religious services, vote, shop, and more. Each of these spheres, realms, or contexts involves, indeed may even be defined by, a distinct set of norms, which governs its various aspects such as roles, expectations, actions, and practices.” (Nissenbaum, 2004, p. 8)

To explain how the concept of contextual integrity could be applied, Nissenbaum posited that contextual integrity is maintained when both “norms of appropriateness” and “norms of flow or distribution” are upheld, and it is violated with either of the norms is violated:

“As the label suggests, norms of appropriateness dictate what information about persons is appropriate, or fitting, to reveal in a particular context. Generally, these norms circumscribe the type or nature of information about various individuals that, within a given context, is allowable, expected, or even demanded to be revealed. In medical contexts, it is appropriate to share details of our physical condition or, more specifically, the patient shares information about his or her physical condition with the physician but not vice versa; among friends we may pour over romantic entanglements (our own and those of others); to the bank or our creditors, we reveal financial information; with
our professors, we discuss our own grades; at work, it is appropriate to discuss work-related goals and the details and quality of performance…In addition to appropriateness, another set of norms govern what I will call flow or distribution of information - movement, or transfer of information from one party to another or others” (Nissenbaum, 2004, p. 8-9).

In an affirmation of the relational focus advanced in by this study, Nissenbaum also noted that the important legal questions of privacy are best addressed in the framework of contextual integrity within which, “it is crucial to know the context - who is gathering the information, who is analyzing it, who is disseminating it and to whom, the nature of the information, the relationships among the various parties, and even larger institutional and social circumstances” (Nissenbaum, 2004, p. 15, emphasis added).

In addition to providing a unifying framework for ethical concerns, maintaining relational integrity is also an avenue to improving the overall quality of the evaluation or research project. Attentiveness to relationship contributes to “interpersonal validity” which is “the trustworthiness of understandings emanating from personal interactions” (Kirkhart, 1995, p. 4). As such, “interpersonal validity inextricably intertwines ethics with trustworthiness” (Marshall & Rossman, 2011, p. 50). In her influential paper “Emerging criteria for quality in qualitative and interpretive research,” Yvonna Lincoln further underscored this point, noting:
“The spiritual, or sacred, side of science emerges from a profound concern for human dignity, justice, and interpersonal respect. The sacredness in the enterprise of science issues from the collaborative and egalitarian aspects of the relationships created in the research-to-action continuum. Researchers who conceive of science in this way make space for the lifeways of others and create relationships that are based not on unequal power, but on mutual respect, granting of dignity, and deep appreciation of the human condition…this dissolution of the hard boundaries between rigor and ethics in turn signals that the new research is a relational research—a research grounded in the recognition and valuing of connectedness between researcher and researched, and between knowledge elites and the societies and communities in which they live and labor” (Lincoln, 1995, p. 284).

Approaching the ethical questions related to visual methodologies through the framework of relational integrity is broadly supported by the literature as both conceptually appropriate and practically helpful. Establishing healthy relationships with participants and maintaining their integrity over time provides a unifying framework for addressing the diversity of ethical questions associated with visual methods, and is a practical means to avoid a well-intentioned but nonetheless damaging ethical paralysis that suppresses the use of valuable visual methodologies.
Focal Research Questions

The goal of this study is to provide a useful supplement and alternative perspective to existing ethical guidelines, but this goal should be understood in a properly modest frame. Ethics is a vast field with conflicting perspectives and values, and providing a ‘definitive moral policy’ related to the ethical use of photos or video is not just beyond the scope of one paper, it is also counter-productive to ethical decision making itself, as such policies “efface thoughtful reflection and can hence nurture orthodoxy and blind absolutism” (Perry & Marion, 2010, p. 99). Working within the conceptual frame of relational integrity, two interrelated research questions were the foci of this study:

1. What are the assumptions and procedures guiding the use of images within the fields of entertainment, journalism, advertising, and the internet?

2. Can broad principles or themes from these procedures be identified to provide a contextually appropriate reference for guiding ethical decisions within the framework of ‘relational integrity’?

Method

The method for this study was developed to address the two research questions and is presented in three categories: design and analytic approach, data sources, and procedure.
Design and Analytic Approach

The primary design approach for this study was content analysis, an approach selected to address the research questions related to the use of images within the fields of entertainment, journalism, advertising, and the internet because it is an established method for “describing and interpreting the written productions of a society or social group” (Marshall & Rossman, 2011, p. 161). In addition to the description and interpretation of the data related to image use, content analysis is a useful method to “identify core consistencies and meanings…The core meanings found through content analysis are often called patterns or themes” (Patton, 2002, p. 453).

The analytic approach within this design was inductive, in that it “involves discovering patterns, themes and categories in one’s data” (Patton, 2002, p. 453). Consistent with established qualitative analysis approaches, this study “prepared and organized the data…for analysis, then reduced the data into themes through a process of coding and condensing the codes” (Creswell, 2007, p. 148) and represented the data in both tables and a discussion. While this analytic process can be described within these sequential and discrete steps, the reality was iterative and emergent - also consistent with best practices in qualitative analysis (e.g., Creswell, 2007, p. 150). The overall validity of the findings was supported by several factors, including the descriptions being context-rich and meaningful, the account being plausible, triangulation between different data sources, linking the data to findings from the literature, and the findings being coherent and unified (Miles & Huberman, 2013, p. 313).
Data Sources

The data for this study comprised existing publicly available information from sources such as sample contracts and image release forms available on the internet, ticket stubs from concerts and sporting events in multiple locations, and publically distributed consent forms gathered through internet searches, personal archives, and requests sent to contacts in Illinois and Colorado. Additionally, documents related to overall intellectual property were gathered and analyzed to provide a legal context for the study. Through this process, a total of 42 documents of varying lengths were gathered, and are marked with a ‘*’ in the references. To facilitate a more specific analysis, the broad category ‘entertainment’ referenced in the title was divided into two sub-categories of ‘film and TV’ and ‘events.’ The number of documents, length, and number of coded excerpts for each category are presented in Table 1 below:

Table 1

Data Categories and Description

<table>
<thead>
<tr>
<th>Source Category</th>
<th>Documents</th>
<th>Average Words / Document</th>
<th>Coded Excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertising</td>
<td>5</td>
<td>276</td>
<td>16</td>
</tr>
<tr>
<td>Internet</td>
<td>9</td>
<td>3,587</td>
<td>47</td>
</tr>
<tr>
<td>Journalism</td>
<td>9</td>
<td>1,246</td>
<td>28</td>
</tr>
<tr>
<td>Film and TV</td>
<td>12</td>
<td>1,000</td>
<td>39</td>
</tr>
<tr>
<td>Events</td>
<td>5</td>
<td>754</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 17 - continued

<table>
<thead>
<tr>
<th>Source Category</th>
<th>Documents</th>
<th>Average Words / Document</th>
<th>Coded Excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Property</td>
<td>2</td>
<td>14,333</td>
<td>9</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>42</strong></td>
<td><strong>2,534</strong></td>
<td><strong>153</strong></td>
</tr>
</tbody>
</table>

**Procedures**

The first step in this study was seeking approval from the Human Subjects Institutional Review Board (HSIRB) from Western Michigan University, and approval was granted in a letter dated January 23, 2013. After collecting and reviewing the documents, relevant excerpts were transcribed or copied into an Excel database to facilitate the coding and data description process (Creswell, 2007). Specialized qualitative analysis software was not utilized for this project because of the relatively small scope of the project and the simple data presentation and ease of use of Excel was considered more useful. While the data analysis process was iterative and more akin to exploring a forest than a sequential process (Creswell, 2007), in order to condense the data and provide a basis for the comparative analysis, the excerpts were coded in what could be described in retrospect as two stages. The first stage of the coding process involved either extracting pertinent ‘in vivo’ codes (Patton, 2002, p. 455) or assigning short summary statements to each excerpt to provide a detailed understanding of the data. The second stage involved reviewing the first stage codes to determine primary themes and categories (Miles & Huberman, 2013, p. 86), and the following five categories were ultimately identified through this process:
a) Who: the individual or group addressed by the agreement (e.g., actor, member of public).

b) Exchange: the nature of what was provided in exchange (if anything) for the image rights (e.g., payment, donation, anonymous use).

c) What: A description of the extent and terms of what was provided (e.g., image, voice, copyright and any specific usage parameters).

d) Territory: The geographic parameters of the agreement.

e) Relationship: A description of the nature of the relationship outlined by the agreement.

Following identification of these categories, the data was reviewed again as each document was coded within the five categories, providing both a unified framework to interpret the data and a higher-level aggregation of the data coded at the excerpt level. Finally, all of the codes from both stages were reviewed an additional time to identify both overall themes and relevant details related to the study questions. The iterative process beginning with developing descriptive open codes to address the research questions led to the clearly defined code categories, and the coded data ultimately closed the loop by informing the findings which addressed the research questions.

Findings

Broadly speaking, the variety of image use procedures from the fields of entertainment, journalism, advertising, and social media further underscored the value
of the relational integrity conceptual frame discussed above, as securing approval for the use of photos or video images in each of these industries is typically a very relational process. Because of this, the ethics of image use within these industries (such as they are!) flow from the factors that inform relationships, such as communication, trust, expectations, and consent. Within the broad framework of relationships, a more specific analysis of the various image use procedures revealed five primary types of relationship related to photo/video images, including:

1. **Recruited Individual:** An image (e.g., personal likeness, image of a location) is legally considered a form of intellectual property. Because of this, a formal exchange of the rights to this property is often necessary, requiring the recruitment of an individual. The terms of this exchange are typically specified in on an image release (for non-professionals) or a property release (for the rights associated with a physical location) or within an employment contract (see below) for those that release images rights as part of their employment. There are two primary motivations for this exchange:
   a) **Financial:** The owner of the image is paid for the image rights. While common in these industries, there are few evaluation contexts where purchasing image rights would likely be appropriate.
   b) **Donation:** Image rights are donated to support a nonprofit cause, for example, to serve as talent in an advertisement for a homeless shelter or a student film project. As many evaluands are non-profit organizations,
seeking image rights as a donation could be appropriate in many evaluation contexts.

2. **Employed Individual:** Film and stage actors, models, TV personalities, sports figures, news anchors, and others are hired to provide services, which inherently require the exchange of image rights, and the parameters of this type of employment relationship are typically established within very specific terms in an employment contract. A related finding from two of the documents was that some people find the glamour and/or recognition associated with being filmed as so desirable that they included language to explicitly prohibit the potential of abuse and/or coercion that is essentially a reverse of employment, stating: “I affirm that neither I, nor anyone acting for me, gave or agreed to give anything of value to any of your employees or any representative of any television station, network or production entity for arranging my appearance on the Picture” (“Group Release,” FilmContracts.net, 2013, “Photographic, video, film, audio recording,” FilmContracts.net, 2013). Again, while this category is common in the industries studied, there are few evaluation contexts where this is likely to be appropriate.

3. **Event Attendee:** Image rights are routinely surrendered (through terms specified in the fine print on tickets, online user agreements, etc.) based on participation in various types of groups. This type of relationship is substantially more distant than the first two above, involving (for example) a person sitting in bleachers and a professional sports team that is filming the event for the arena jumbotron and/or
TV broadcast. The types of group participants identified through this analysis include:

a) Attendee of entertainment events (e.g., shows, sporting events, and theme parks)

b) Participant in a live television audience (e.g., game show or talk show)

c) Internet User (e.g., someone who uploads video/photos on social media sites)

While not likely common, potential evaluation contexts involving attendees of events could include participants in workshops or large scale training courses, community events, or museum or zoo event attendees, for example.

4. Anonymous Public: This category is similar to ‘event attendee’ above, but broader as it relates to gathering images in public, free, or uncontrolled venues, to be distinguished from paid events. While individuals might be recognizable in these situations, they are effectively treated as anonymous, or ‘nameless’ (Merriam Webster, 2013) because either they are in the background of the primary subject or one individual in a shot of a crowd. In these situations there is typically little to no formal ‘relationship’ expressed through notification, legal forms, or explicit informed consent. Examples of the filming situations falling within this category and typical methods of communication with filmed or photographed subjects in these situations include:
a) Films or photography taken in public locations (e.g., on a city street, in a subway station). A sign notifying people with image release language is sometimes posted in these situations.

b) News events (e.g., potentially by providing an interview, but more likely being in the background of a newscast). News producers often try to secure written approval when possible (especially of interviewed individuals), but in many crowd settings this is often not possible, and is not legally required if filming is conducted in a public area without the reasonable expectation of privacy. The freedom to use images for news or educational purposes is covered in the ‘fair use’ clause of copyright law (United States Copyright Office, 2012) and is also the legal justification for the photographs or video taken of celebrities in public places by paparazzi (Mr. Paparazi.com, 2013). As illustration, if movie stars are in a public area without ‘reasonable expectation of privacy,’ the fair use clause is such that they cannot prevent a photo/video from being used as a tabloid ‘news’ story, but that same image would be restricted from use an advertisement or endorsement.

c) Security cameras. The growing reality of video surveillance raises complex and important ethical questions addressed in other articles (e.g. Nissenbaum, 2004), but the fact remains that security cameras are gathering visual data in an increasing number public venues across the world. For example, in a remarkable irony, there are now 32 security cameras within 200 yards of the London flat where George Orwell wrote the classic book ‘1984’ about a
repressive ‘Big Brother’ that is always watching (London Evening Standard, 2007).

Various evaluation contexts could utilize visual methodologies involving the ‘anonymous public’ in similar ways, such as utilizing security camera data for basic attendance or demographic information.

5. **Community Participant:** This final category is based on the unique aspects of the image usage agreements for the social media and web services, including Facebook, websites owned by Google (e.g., YouTube and Picasa), Pinterest, Tumblr, and Flickr. In each of these examples, the relationship between the companies and the ‘participants’ or users of the services related to visual data involves a combination of exchange and group participation:

   a. **Exchange:** While specific wording or structure varied, all agreements for these social media sites involved an element of exchange, whereby the user is receives free use to the servers, software and service of the website, and grants to the website a non-exclusive, worldwide license to the words, images, and video they uploaded. The first reason for this structure is that the websites require the license to the content to legally operate; they couldn’t host the images and serve them to other users around the globe without it. The second and related reason is that this release gives the companies the legal means to generate revenue; for example, it allows them to analyze the uploaded content to more effectively target their ads.
b. **Group Participation**: A primary value of social media sites is that they comprise a community whereby individuals receive benefits for participation, such as by being able to enjoy new means of communication, receive feedback on their photographs from a global community, or gain creative inspiration, to name but a few such benefits. These communities function within an established social contract that is similar to offline group activities (like a ‘potluck’ meal). The structure of these virtual communities involves individuals voluntarily surrendering their individual rights or property (i.e., a non-exclusive license to their photographs) to be able to benefit from participation in a community with others that have done the same. For example, since everyone uploading photos to Facebook grants a non-exclusive license to their photos, each individual gains the advantages of being part of a community whereby photos are easily shared with family and friends around the world - for free.

This community participant category is perhaps the most relevant to evaluation contexts and the related applications will be discussed further below.

**Discussion**

The potential scope and level of relationship for gathering images in evaluation and research projects spans a similar range to the five types noted in the findings above, and these categories could be useful reference for guiding ‘relational integrity’ decisions in the context of specific evaluation projects. Furthermore, just as the findings indicated that methods of informed consent vary with the type of relationship...
within these other fields, evaluators, researchers, and IRB committees can use these categories as reference in determining the most contextually-appropriate approach to informed consent for varying projects. The following examples of the different approaches to relationship within the different categories are offered as illustration:

**Recruited Individual** – Videotaping or photographing volunteers requires the time intensive process of identifying and recruiting individuals, which could be appropriate in evaluations involving especially deep or long-term involvement. In every industry studied, a similar relationship would involve a signed and specific release form related to the exchange of intellectual property, typically involving the transfer of copyright ownership to facilitate future use.

**Employed Individual** – While unlikely in most evaluation or research scenarios, it is possible that University employees (or graduate student assistants) could provide visual data as part of their employment (e.g., regular video blog, workplace studies) and in every industry these situations require including the scope and terms of this intellectual property exchange in the employment contract.

**Event Attendee** - Various events or group activities are recorded in research or evaluation projects such as conferences, workshops, or recipients of nonprofit services. In most of these contexts, it is not feasible to develop relationships or secure informed consent releases from each individual in a group. This study found that in similar group event situations, participants are typically informed of recording through existing, non-intrusive means such as signage, fine print on tickets, or announcements
at the beginning of an event – all of which could be feasible approaches within evaluation contexts.

Anonymous Public - People in public venues (streets, parks, etc.) photographed or filmed at wide angles could be recorded for research and evaluation studies. Within the image use policies reviewed for this study, most involved very low levels of communication related to this type of public recording – either none at all to a simple sign. One example involving the anonymous public included an evaluator seeking to estimate the number of homeless people in a community who filmed public areas such as street corners or sidewalks from a van driving a specific route (Berry, 2007).

Community Participant – Many research and evaluation projects are explicitly developed to benefit a specific community or organization, such that participation in these projects benefits both individual participants and a larger group. Just as individuals might be willing (or even excited) to participate in a drug trial because they want to be part of finding a cure for both themselves and a wider community, so too individuals could be enthusiastic about sharing their photos, video or other visual data for a project that will clearly benefit a community or issue they care about (Clark, 2010). Projects with abstract or ‘ivory tower’ goals are unlikely to meet this criteria, so honest appraisal of the benefits from the perspective of the participants is essential. However, if the project provides clear benefit for participants and/or a community or issue, the spirit of relational integrity would involve approaching the involvement of participants as potential beneficiaries of a community project, in contrast to an apologetic request to subjects for presumably intrusive data. Therefore, the time
participants give to participate and the release of rights to visual data can be understood within a framework that is similar to participation in social media sites, whereby receiving the social and personal benefits of participation is exchanged for image rights which are released by all members of the group equally. On this basis, it is a violation of this type of relationship and project to frame the participants as ‘subjects’ that need protection from a researcher that is extracting data, yet this is the implication of some IRB processes (Allen, 2009) and is a particular risk for those who conceive of their primary ethical concern in terms of avoiding legal liability for their institutions (Wiles et al., 2008) rather than respecting the relationship of participants (Lincoln & Tierney, 2004).

The Culture of Research and Evaluation

In addition to these five types of relationship developed from the data analysis, the first focal research question prompted a deeper reflection on the assumptions behind the structure of the data and procedures. Since these assumptions are implicit and not able to be coded and analyzed with deductive reasoning, this reflection was inductive and led to broader conclusions. First, the fields referenced are very large and maintain a position of cultural leadership, unlike the small and largely marginalized use of visual methods in the research and evaluation communities. Similarly, in most industries referenced, images are the primary communication method and words and numbers serve a supporting role (e.g., football games with scores/stats, feature films with subtitles, commercials with closing title slides, internet memes with dominant
images and short text, etc.) while the reverse is the case in most research and evaluation communications. The assumption framing much of the work in these industries is that the opportunity to be filmed or recorded is inherently desirable and generally contributes to social prestige for participants (e.g., be on TV, projected on stadium jumbotron), while the framing within the research and evaluation community is often the opposite, assuming that recording of images involves a deeper degree of sacrifice and intrusion. Taken together, these differences point to a considerable cultural divide related to recorded images between the more scholarly culture of researchers and evaluators and the broader public; those seeking effective communications and relationship development should therefore consider the cross-cultural implications of reaching out of the research and evaluation sub-culture. While these reflections were guided by the first research question, the findings address the second. While delimited as an initial exploration, both the specific findings and broader reflections of this study did address the research questions by identifying broad principles and themes that can serve as contextually appropriate guidance in what amounts to cross-cultural engagements surrounding visual methodologies.

Implications for Informed Consent Practices

The goal of the informed consent process is to cultivate and maintain a healthy, appropriate relationship between evaluator or researcher and study participants. While this goal is conceptually simple, implementation involves considerable complexities and operates within limitations. Because of this, it should be understood that there is
no form or process that covers every eventuality; ultimately informed consent is one part of a fluid/evolving relationship between researcher and participant. Furthermore, just because someone signs a release doesn’t mean they understand it or will change their minds later, so being mindful of the evolving relationship is critical (Wiles & Crow, 2007, Fang & Ellwein, 1990).

a) Consent forms should be understood as a tool to deliberately and thoughtfully build relationships that ensure that the parameters of the image data collection and use are understood and agreed – more an introduction than final contract. Admittedly this ideal is difficult to implement, as the thoughtful development of relationships over time operates in tension with the pressures of project schedules, budgets, and concern to limit legal liability.

b) The complexity of the Internet is such that terms and conditions of internet companies are very long and written by lawyers and for lawyers - not participants. Some websites (e.g. Pinterest, 2013 and Flickr, 2013) have addressed the relational barrier this creates by providing short, conversational statements after each block of full legal text. This approach could be considered as a practical (though not ideal) solution for researchers and evaluators, as it provides both the completeness required by IRB or legal departments and also facilitates relational connection. An even better solution might be to only provide participants the relationally connective ‘plain English’ language with a web link to the full legal text for the rare few that might want to read it.
c) If the request for participation in a research or evaluation project is going to be intrusive or potentially objectionable, this should be communicated from the beginning in both the discussion and consent form. For reference, in the film industry filming of nudity or sex acts is agreed in advance on an addendum to standard image releases to ensure clear approval and consent of this most personal type of visual data (Film Contracts “Nudity Rider and Casting Agreement,” 2013).

d) The method of informed consent should be kept within the parameters of the level of relationship. For example, the huge crowds attending an NFL games do not sign separate ‘image releases’ in the event that they might be recognizable on TV; the simple statement on the fine print on the ticket has been generally accepted as appropriate for that context (National Football League, 2011). Similarly, researchers and IRB committees should explore varied methods and approaches to informed consent to realistically and appropriately honor the relationship for different research projects and not insist on a single template for all projects and levels of relationship, especially since “written documentation of informed consent can create harm for research participants in some circumstances” (American Journal of Evaluation, 2005).

Conclusion

In addition to providing rich and useful qualitative data in research and evaluation projects, photos and video are a form of communication that can be uniquely powerful and relevant for shaping ethical decisions and behavior. While
“photographs cannot create a moral position…they can reinforce one – and can help build a nascent one” (Sontag, 1973, p. 17). As just one example of a photograph reinforcing or building a nascent moral position, “a single photograph may have done more to halt the Vietnam War than all the writings of moral philosophers of the time put together” (Blackburn, 2002, p. 5). The photo being referenced is of Kim Phuc as a child running naked down a street after a napalm attack, and the iconic ‘napalm girl’ photo is among the most famous of all time (KimFoundation.com, 2013). In considering the overall ethical framework for visual methodologies, the ethical implications of either facilitating or restricting this powerful form of communication must be considered in addition to the rights and relationship with participants. Words and numbers are a relevant and effective means of communication in many situations, but photos, video and other visual means fill a unique and growing role, and it is inappropriate at best and unethical at worst for visual methodologies to be marginalized or required to fit ethical approaches developed for words and numbers. To address this need, the conceptual framework of ‘relational integrity’ offers a useful approach that compliments the practices of both existing ethical Institutional Review Boards and the guidelines provided by the International Visual Sociology Association (Papademas, 2009). Within this context, the image use procedures from the fields of entertainment, journalism, advertising and the social media provide contextually relevant guidance for informing the relationships and ethical decisions of researchers and evaluators utilizing visual methodologies.
V. CONCLUSION

Overview

The problem addressed by this dissertation is the underutilization of visual methods within the field of evaluation, and the associated overarching goal was to address barriers to using this important source of data. The specific barriers addressed include the learning curve associated with technology (Travers, 2009), longer time needed for working with visual data (Fang & Ellwein, 1990), cost (Heath, Hindmarsh, & Luff, 2010) and concern over the uniquely challenging ethical questions associated with visual methods (i.e., Allen, 2009; Wiles et al., 2008; Emmison & Smith, 2000; Perry & Marion, 2010). These barriers were addressed through three interconnected papers: the first two addressed barriers related to technology, cost, and time and the third addressed barriers related to ethical questions.

Papers one and two describe exploratory studies of new analysis methodologies utilizing the (relatively) user-friendly, low cost, and time-efficient Mturk crowd-source platform to code user-generated product review videos. Paper one focused on the open-coding stage of analysis and paper 2 addressed the categorization of the video data within pre-determined codes. The third paper presents a conceptual framework for addressing unique ethical challenges posed by visual methods and offering specific suggestions for ethical use of visual data based on the image use procedures from the domains of entertainment, advertising, journalism, and the internet.
Paper 1 Summary

Conclusions

The findings indicated that Mturk participants provide more analytic data than required by the internet survey instrument used to facilitate their participation, and provided both detailed descriptive codes and candid overall impressions of the video data. A primary quantitative finding in the comparison between the Mturk participants and evaluator control group was that the average words provided per second of video data decreased for the evaluators over the course of the five videos they reviewed, while on average the Mturk participants provided relatively consistent quantity throughout. A primary qualitative finding in this comparison was that the evaluator coding and descriptions were more broad, professional and evaluative, while the Mturk participants were more detailed, descriptive, and raw or sometimes crass. These findings suggest specific limitations to this methodology (noted below).

However, utilizing Mturk participants for the open coding of video data offers considerable potential advantages for evaluators: the broader perspective available through a diverse group of participants provides a uniquely useful source for relevant analytic perspective that is both time efficient and cost effective.

Limitations

This study utilized user-generated product review videos as the primary source of data, so this comparative methodology study is delimited within that context. Similarly, this was a small scale pilot-study, so the findings are limited to this single
point in time, by the small dataset, and small group of participants. The methodology was also limited in that neither the selection of Mturk participants nor comparison sample was random, and therefore selection bias is conceptually a concern; Mturk participants were qualified with a basic level of experience and proficiency (500 HITS completed and 90% approval) and living in the United States, and the qualitative analysis control group was the primary researcher plus two fellow students in the Western Michigan University Interdisciplinary PhD in Evaluation program. That said, Mturk HITS commonly require similar qualifications and workers are free to login at any time, so the opportunity to participate was considered as broad as typical for Mturk and ultimately more related to the effectively random chance of logging in at the time the surveys were released. Likewise, the selection of the control group followed a similar process as typical qualitative research studies, recruiting experienced and trained individuals with availability in their schedule. The different size of the groups in the comparison is another limitation of the methodology (3 in control vs. 25 Mturk participants per survey) but this limitation was inherently a basis for the comparison itself, as qualitative analysis coding is typically completed by 1-3 evaluators because of cost and time constraints not operative using the Mturk platform. Furthermore, this limitation was addressed by calculating overall group averages and analysis at the group level, not sums of individual responses. None of the limitations were considered serious threats to the validity of the overall findings, though they are factors to consider as context for interpretation. Future opportunities to address many of these limitations will be noted in the ‘future research’ section below.
While one goal of this study was to overcome the barriers to use of visual methods, it is recognized that the Mturk platform has various limitations and could therefore present new barriers even as it seeks to address others, for example:

a. Mturk is a software platform that is unfamiliar to most evaluators and researchers, such that it requires overcoming any fears and negotiating a learning curve associated with new technologies.

b. The methodology described involved linking the Mturk participants to video data available on YouTube, and some evaluation projects would require that the data be confidential. Other video serving platforms could address some of these concerns, but the use of Mturk participants in an online environment inherently involves less control of the data than current standard practices.

c. While this methodology was considered low cost and time efficient, it is also recognized that it is neither free nor instant, such that potential barriers related to cost and schedule could remain.

d. Mturk is a cloud-based service which requires a reliable internet connection, so will not be appropriate in all geographic contexts.

e. Some have raised ethical questions related to ‘digital sweatshop’ potential related to the low hourly wage received by most Mturk workers (Scholz, 2012, p. vii), such that the ethical concerns must be considered in each situation.
Future Research

As a pilot test, the primary opportunity for future research is to replicate the study using more data, a varying number of Mturk participants and over a longer period of time to test and refine the methodology within other contexts. Similarly, additional research to replicate the methodology using different data (i.e. photographs, interview footage) could explore the potential of utilizing Mturk participants for qualitative analysis more broadly.

Paper 2 Summary

Conclusions

The findings of this exploratory study indicated that crowd-sourced workers can provide very accurate coding within specific parameters. For example, the coding for gender showed perfect or nearly perfect accuracy across all coding cohorts, indicating that objectively clear and simple coding can be provided accurately with even very small coding cohorts. Similarly, the very high accuracy of coding for video quality of the radio controlled toy (equaling the accuracy of gender) indicates that crowd-sourced workers can also provide accurate and consistent coding even for more subjective decisions. That said, the high variability and generally low accuracy in coding for video quality for the other videos indicates that more detailed categorization formation, comparison images, or other preparation materials should be considered to achieve high accuracy on a wider range of videos. The generally high
accuracy of coding product features which were present and other ‘positive cases’ indicates that crowd-sourced workers can identify specific content details.

Overall this study contributed to the overall goals of this dissertation by demonstrating that crowd-sourcing can effectively address two of the primary concerns related to the use of video data – time and cost. As relates to time, while there is additional work associated with learning and developing the tools on the SurveyMonkey and Mturk platforms, once those initial steps are completed, the actual coding was completed very quickly by the Mturk workers; within 15 minutes for the 3 coder cohort and within 4 hours for the 10 and 20 coder cohorts. As for cost, the 165 completed surveys (or HITS) cost a total of $117.98 (Mturk participants plus Amazon commissions), or $0.55 for the shorter videos and an average of $0.715 per completed survey.

Limitations

This study was designed to explore the viability of crowd-sourced categorical coding as a compliment to a similar study exploring crowd-source open coding, (Wilson & Azzam, Manuscript in Preparation) and as such must be understood in the context of several important limitations. First, the use of the user generated product review videos for both studies is such that other types of video data relevant to evaluators (i.e. interviews, classroom interactions, training sessions, etc.) were not addressed. Similarly, this study was designed to explore a diversity of categorical coding tasks, and as such provided initial indicators across a broader range of
situations as opposed to strong evidence of specific tasks. Like paper 1, this was a small scale pilot-study, so the findings are limited by the single point in time, small dataset, and relatively small group of participants. Additionally, the selection of Mturk participants was not random, and therefore selection bias is conceptually a concern; Mturk participants were qualified with a basic level of experience and proficiency (500 HITS completed and 90% approval) and living in the United States. However, as with Paper 1, Mturk HITS often require similar qualifications and workers are free to login at any time, so the opportunity to participate was considered as broad as typical for Mturk and very nearly random as the 5 surveys were released at different times in the day to maximize variation.

**Future Research**

Additional research is needed to provide more rigorous and extensive testing of the methods outlined in this initial exploration and to identify the contexts and procedures to generate broadly applicable, stable and accurate coding. In specific:

1. Coding for gender was demonstrated as highly accurate in all three cohorts, and future studies could test the reliability of other basic demographic or contextual factors such as race, age category, or setting.
2. The video with the clearest picture quality (radio controlled toy) was coded with very high accuracy in all cohorts, but the other videos demonstrating varying levels of camera quality, lighting and video compression were highly variable, and future
research could test providing comparison images, worker training or screening, or a wider range of categories related to audio or video quality.

3. The coding related to personal judgments of emotion (trustworthiness of the presenter, overall emotional impact) showed very little evidence of bias and general correspondence with the open coding study (Wilson & Azzam, manuscript in preparation). Therefore, specific and rigorous tests of the reliability of emotive coding should be conducted.

4. The coding related to details of the video content (which generally comprises the core goal of video analysis projects) was highly variable across the different videos, cohorts and coding questions. Additional research should identify the number of code categories, code descriptions, training and preparation of coders, and other design features needed to consistently produce accurate coding of the details of video data content.

Paper 3 Summary

Conclusions

The theme of relationships is prevalent throughout the visual ethics literature, and indeed through ethics literature in general, and this theme was utilized as the basis for the conceptual framework of ‘relational integrity’ that is suggested as useful for approaching visual ethics questions. This conceptual frame was supplemented by the findings from a comparative analysis of the image use procedures in the fields of entertainment, journalism, advertising, and social media, which found five different
categories of relationship and associated relational structures. The five categories of relationship include recruited individual, employed individual, event attendee, anonymous public, and community participant. The suggested applications of the ‘relational integrity’ conceptual framework and findings from the comparative analysis include guidance on how to develop and maintain consent forms within an overall relationship-building context, ensuring that any potentially objectionable research needs are communicated in writing in advance, and keeping informed consent in the parameters appropriate with the level of relationship.

Limitations

The focus of both the literature reviewed and data analyzed is within the very narrow domain of visual methodologies and image use procedures, and this study is explicitly delimited within that context.

It is recognized that each of the industries included in the comparative analysis is rife with unethical practice related to image use, so those sources are explicitly limited in the ethical guidance they provide. The findings from the comparative analysis are presented as relevant contextual guidance and as a framework for approaching the ethical questions of research and evaluation projects, but should not be used as sources for ethical norms.

Even though much of the literature reviewed was from the broader social sciences and many of the conclusions could have direct application in a wide variety of studies employing visual methods, this study is a component of an evaluation
dissertation, and should therefore be understood as especially suited to the perspective and needs of evaluators; potential applications within the broader social science research community requires additional research as noted below.

**Future Research**

The first opportunity for additional research relates to the last limitation noted, and that is to confirm that the findings of the comparative analysis and ‘relational integrity’ conceptual frame are relevant within the broader social science research community. Since most of the literature reviewed was from this wider community (i.e. not specifically related to evaluation) and many evaluation methodologies were derived from the social sciences (Worthen, Sanders, & Fitzpatrick, 1997) the needed adjustments might be slight, but review and feedback from 2-3 social science researchers could help in that regard.

The second opportunity is to explore how the ‘relational integrity’ conceptual frame could be relevant and useful for approaching a wide variety of ethical questions, not just those using visual methodologies. Prioritization of relationship with participants is a practical and conceptually promising path toward implementing the ‘golden rule’ of “do unto others as you would have them do unto you” (Matthew 7:12 King James Version) that is implicit in many ethical approaches beyond the Christian source of the text, so further work to explore the broader application could be valuable.
Contribution to Evaluation

Contribution to Theory

This dissertation made an initial contribution toward charting the “great unexplored frontier” of evaluation under conditions of complexity (Patton, 2011, p. 1) by focusing on the feedback mechanisms that are a central feature of complexity (Wheatley, 1999). While Developmental Evaluation (Patton, 2011) describes how an evaluator can provide valuable perspective and service to the ongoing evolution of a complex program, and systems thinking (i.e., Hargreaves & Podems, 2012; Williams, 2008) provides a good conceptual framework for approaching complex situations, no evaluation literature focused specifically on feedback mechanisms, such that this contribution is unique.

Papers 1 and 2 utilized user-generated product reviews, and therefore contributed to literature related to user-generated evaluation (Scriven, 2013). While this contribution was largely related to analysis methodology (further described below), it was also a contribution to the theory, because it argued that ‘the crowd’ is a valid and useful source for the interpretation and analysis of the feedback provided by ‘the crowd.’

Paper 3 contributed to the theory of visual ethics in two ways: First, it built on the theme of ‘valuing relationship with participants’ found throughout the visual ethics literature and suggested that ‘relational integrity’ is an appropriate and useful conceptual frame for addressing ethical questions. Second, it outlined the four categories of relationship found in an analysis of image use procedures in the fields of
entertainment, advertising, media and the internet. These two aspects of the theoretical contribution were unified in the closing presentation of specific implications for research and evaluation projects using the ‘relational integrity’ framework within the four categories of relationship found in the image use procedures.

**Contribution to Methodology and Practice**

The intended contributions to the practice of evaluation for the three studies completed as part of this dissertation are noted in the summary of each paper above. Papers 1 and 2 pilot tested new analytic methods for video data that can be implemented in the ‘real world’ of evaluation practice because they directly address two of the three concerns noted in Real World Evaluation – time and cost. (Bamberger, Rugh, & Mabry, 2006). Toward that point, this study certainly required time investment for the initial development of this approach, from setting up the accounts, developing the online surveys with embedded videos, etc. However, once those initial steps were completed, the analytic data from Mturk was available within 12 hours, as compared to the week between the initial request and the receipt of data from the evaluation team. As for cost, the 125 surveys cost a total of $74.52 (Mturk participants plus Amazon commissions) for an average of $.596 per completed survey.

Additionally, these methods improved the overall rigor or value of the analysis process by providing a realistic method to increase both the quantity and conceptual quality of the perspectives on the analysis:

140
1) Quantity: The typical qualitative analysis process which was utilized by the evaluator comparison group involved a researcher coordinating a small research team to complete the coding, such that the total number of perspectives represented was 3. The methodology utilizing Mturk that was pilot tested and described in papers 1 and 2 involved a researcher and 25 Mturk participants per survey, such that this method provided exponentially more perspectives on the video data.

2) Conceptual ‘Quality’: The typical process addresses the limitations of perspective by the researcher and/or research team through a combination of memos but especially through member checks (Patton, 2002; Seale, 2002) of the completed analysis. The methodologies using Mturk build on the conceptual value provided by ‘member checks’ by essentially having the ‘members’ (defined as a cross-section of people in the United States) directly participate in the analysis. Conceptually this provides insights that are more authentic within that community because it is a more direct path to the longstanding principle of qualitative research which is that “the whole analysis of experience must be based on their concepts, not ours” (Boas, 1943, p. 314), or for the analyst to get “out of the way of the data to let the data tell their own story” (Patton, 2002, p. 457). The Mturk platform provides a practical method for members of “the community” to directly provide their own conceptual frame and perspective on the raw data itself…not post-facto ‘checking’ of the conclusions after the researchers impose the frame developed with the intention of representing the community.
REFERENCES


10.1177/0193841X06296947


10.1002/mar.20427


*Directors Guild of America (2013). Director Deal Memo. (DGASigPkFrm: 08AB)


http://www.expotv.com/faqs#What-is-EXPO

*ExpoTV Video (2013). Apple 8 GB iPod.* Retrieved 7/02/13 from:
http://www.youtube.com/watch?v=V3E3vexk39Y

*ExpoTV Video (2013). Radio Controlled Toy Helicopter.* Retrieved 7/02/13 from:
http://www.youtube.com/watch?v=AGNew9RmwuM

*ExpoTV Video (2013). Samsung A900 Cell Phone.* Retrieved 7/02/13 from:
http://www.youtube.com/watch?v=0BtaCcjolxY

*ExpoTV Video (2013). Settlers of Catan Game.* Retrieved 7/02/13 from:
http://www.youtube.com/watch?v=Hvo_i1vdDEo

*ExpoTV Video (2013). St. Ives Facial Moisturizer.* Retrieved 7/02/13 from:
http://www.youtube.com/watch?v=7SvadWmaHvQ

http://www.facebook.com/legal/terms


doi:10.1177/1468794111436140


*Film Contracts (2013). Talent use of name & likeness in a film or TV clip. [PDF Document]. Retrieved from:
http://www.filmcontracts.net/contracts/form.php?id=1181


http://www.flickr.com/guidelines.gne


http://www.popularmechanics.com/technology/digital/visual-effects/a-brief-history-of-the-movie-camera#slide-1


Downloaded 5/27/13 from:


http://visualsociology.org/journal.html


Kaiser Permanente (2012). *Getting Started in Video Ethnography – A Catalyst for Guiding and Motivating Quality Improvement*. Retrieved 12/05/12 from:


*Evaluation Practice. 6*(1), 1-12.


http://www.merriam-webster.com/dictionary/anonymous?show=0&t=1372181365


*Mr. Paparazi (2013). Legal Terms. [Web Document]. Retrieved from:

https://requester.mturk.com/tour


http://www.michaelscriven.info/images/ULTIMATE_EVALUATIONQUESTIONS.pdf


http://ww.telegraph.co.uk/technology/social-media/9780565/Facebook-terms-and-conditions-why-you-dont-own-your-online-life.html

155


156


APPENDICES

A: Paper 1 Survey Sample

1. Describe the video content by entering 3 - 5 words or short phrases for each :30 second segment of video:

   0 - 30 Description 1
   0 - 30 Description 2
   0 - 30 Description 3
   0 - 30 Description 4
   0 - 30 Description 5
   :30 - 1:00 Description 1
   :30 - 1:00 Description 2
   :30 - 1:00 Description 3
   :30 - 1:00 Description 4
   :30 - 1:00 Description 5

2. Please provide a narrative description of the video & the impact it had on you - as if you were telling a friend.
3. What is your gender?
- Female
- Male

4. Which category below includes your age?
- 17 or younger
- 18-20
- 21-29
- 30-39
- 40-49
- 50-59
- 60 or older

5. What is the highest level of education you have completed?
- High school graduate
- Some college
- Trade/technical/vocational training
- College graduate
- Some postgraduate work
- Postgraduate degree

6. Which of the following categories best describes your employment status?
- Employed, working 1-39 hours per week
- Employed, working 40 or more hours per week
- Not employed, looking for work
- Not employed, NOT looking for work
- Retired
- Disabled, not able to work
7. What is your marital status?
   - Single/never been married
   - Married
   - Separated
   - Divorced
   - Widowed

8. What is your race? Please choose one or more.
   - Native Hawaiian or Other Pacific Islander
   - White
   - American Indian or Alaska Native
   - Hispanic or Latino
   - Black or African American
   - Asian
   - Other

9. Please enter the two letter abbreviation for the state you live in.

Please enter the following completion code on Mturk: ExpoTV/Samsung
1. Was the product reviewed something you personally might use?
- Yes
- No

2. What is the gender of the presenter?
- Female
- Male

3. What identifying information about the product was provided?
- Model only
- Manufacturing company only
- Model and manufacturing company

4. What features of the product were described?
- Strengths only
- Weaknesses only
- Strengths and weaknesses
### ExpoTV Categorization

5. **What product features were mentioned (select all that apply):**
   - [ ] Touchscreen
   - [ ] Internet connection
   - [ ] Mp3 Music player

6. **What product features were mentioned (select all that apply):**
   - [ ] Maps
   - [ ] Apps
   - [ ] Telephone

7. **What was the quality of the video picture?**
   - [ ] Blurry
   - [ ] Clear

8. **Did the reviewer verbally recommend the product?**
   - [ ] No verbal recommendation
   - [ ] Mixed verbal recommendation (i.e. “It’s o.k.” or similar)
   - [ ] Strong verbal recommendation (i.e. “I definitely recommend” or similar)
   - [ ] If yes, what was said to recommend the product?

9. **What emotional impact did the review have on you?**
   - [ ] Strongly negative
   - [ ] Somewhat negative
   - [ ] Neutral
   - [ ] Somewhat positive
   - [ ] Strongly positive

10. **How trustworthy was the reviewer?**
    - [ ] Not trustworthy
    - [ ] Somewhat trustworthy
    - [ ] Very trustworthy
11. What is your gender?
   - Female
   - Male

12. Which category below includes your age?
   - 17 or younger
   - 18-20
   - 21-29
   - 30-39
   - 40-49
   - 50-59
   - 60 or older

13. What is the highest level of education you have completed?
   - High school graduate
   - Some college
   - Trade/technical/vocational training
   - College graduate
   - Some postgraduate work
   - Postgraduate degree

14. Which of the following categories best describes your employment status?
   - Employed, working 1-39 hours per week
   - Employed, working 40 or more hours per week
   - Not employed, looking for work
   - Not employed, NOT looking for work
   - Retired
   - Disabled, not able to work
15. What is your marital status?
- Single/never been married
- Married
- Separated
- Divorced
- Widowed

16. What is your race? Please choose one or more.
- American Indian or Alaska Native
- Native Hawaiian or Other Pacific Islander
- White
- Asian
- Hispanic or Latino
- Black or African American
- Other

17. Please enter the two letter abbreviation for the state you live in.

Please enter the following completion code on MTurk: ApplePodExpoTV2
Date: January 23, 2013

To: Chris Coryn, Principal Investigator
Kurt Wilson, Student Investigator for Dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed for HSIRB Project Number 13-01-30

This letter will serve as confirmation that your project “The Ethics of Image Use: A Descriptive Analysis of Image Use Procedures in the Fields of Entertainment, Journalism, Advertising and the Internet” 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 and not collecting personal identifiable (private) information about individuals.

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.
Date: June 6, 2013

To: Chris Coryn, Principal Investigator
    Kurt Wilson, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 13-06-15

This letter will serve as confirmation that your research project titled “Crowd-Sourced Qualitative Data Categorization: New Tradeoffs in Speed, Cost and Rigor Utilizing the Mark Platform” has been approved under the exempt category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: June 6, 2014
Date: June 24, 2013

To: Chris Coryn, Principal Investigator
    Kurt Wilson, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 13-06-29

This letter will serve as confirmation that your research project titled "Crowd-Sourced Open Coding: Harnessing Diverse Perspectives to Analyze Rich Qualitative Data Using the "Mturk" Platform" has been approved under the exempt category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: June 24, 2014