FUZZIER CONSTRUCTS IN CONTENT ANALYSIS: RATINGS FROM THE CROWD VS. TRADITIONAL EXPERTS

by

Miles Moffit

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Arts in
Industrial & Organizational Psychology

Charlotte

2018

Approved by:

_____
Dr. Steven Rogelberg

_____
Dr. Clifton Scott

_____
Dr. Eric Heggestad

©2020
Miles M. Moffit
ALL RIGHTS RESERVED

ABSTRACT

MILES MOFFIT. Fuzzier constructs in content analysis: Ratings from the crowd vs. traditional experts. (Under the direction of DR. STEVEN G. ROGELBERG)

The task of rating (or "coding") text data as a component of qualitative content analysis remains a time-consuming process. Advances in crowdsourcing platforms have presented novel opportunities to outsource this work away from the desks of academics, graduate students, and other "traditional" Subject Matter Experts. Recent studies suggest that the crowd could produce reliable ratings of qualitative constructs more exclusive to the social and organizational sciences, which could dramatically alter the way content analyses are conducted, largely by reducing the time requirements of those analyses. This is particularly true for what might be called "fuzzy" constructs; constructs without logical boundaries and traditionally considered less accessible to non-specialist raters. This study makes use of extant fuzzy construct data from an archival study (n = 177), ratings of subject matter experts (SMEs) from the same study (n = 6), and newly collected crowd ratings (n = 96) of the same data rated by the SMEs. Comparisons of the crowd's ratings relative to the graduate-level researchers' (i.e. "traditional") ratings revealed a high level of similarity. Crowd-based groups of as few as six randomly selected raters were similarly reliable to groups of three traditional experts, while not significantly more or less accurate. When specific selection criteria were used instead of random selection, as few as three crowd-based raters were likewise similarly reliable and accurate. These and other data on hand support a set of future recommendations for qualitative researchers, as well as for further empirical investigation into the use of crowd-based raters for traditional research and content ratings tasks.

ACKNOWLEDGEMENTS

Like many projects of its kind, this document is the culmination of a series of incremental additions, revisions, rethinkings, and re-imaginings.

As of the moment of this writing, a significantly shortened version, edited for publication in particular journals in particular fields, is in the process of being considered for publication. It is fitting, however, that the enclosed version of the document is both the one I defended before my committee as my thesis project, and the longest version of the paper. As the enclosed represents what became a significantly grander, longer, more challenging undertaking than was first envisioned, it was my thinking that to submit any subsequent, shorter version—though there have been many—would not present the entire content of the project as was completed and defended.

My deepest thanks to my first mentor and chair, Dr. Steven G. Rogelberg, for making the suggestions and asking the questions that led to the project, for his detailed, nuanced guidance, for his enduring patience and understanding, and most of all for having faith in me as his protégé and as his student. My gratitude extends as well to my current mentor, Dr. Clifton Scott, whose counsel and encouragement has been a constant source of inspiration and affirmation in equal measure, five years running. And to Dr. Eric Heggestad, whose methodological expertise afforded me valuable insights into my analytical approach and to my results that I might not otherwise have considered in the development of the paper.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The coding of text-based data for content analyses, whether having been collected from research participants in a laboratory or from the public domain, is an essential practice in much of qualitative and quantitative research. First practiced following the invention of the printing press, content analysis—a set of practices for analyzing and describing written communications in terms of latent themes, patterns, or biases (Holsti, 1969)—and its various iterations have been under continual refinement. What began more than two centuries ago as a means to detect anti-establishment language in early newspapers (Groth, 1948; as cited in Krippendorff, 2013) now exists in the form of lightning-fast computer algorithms that can analyze a body of written words at the push of a button.

Among the more recent innovations of technological surge, "crowdsourcing" (Howe, 2008) has emerged in the last decade as a more cost-effective and efficient data collection technique. The term "crowdsourcing" was coined in reference to the outsourcing of tasks traditionally reserved for company employees or subject matter experts (SMEs) to a larger group of non-experts with the objective of generating similar results to those of the traditional group (Howe, 2006). A December, 2014 White House blog post defined it as "a process in which individuals or organizations submit an open call for voluntary contributions from a large group of unknown individuals ('the crowd') or, in some cases, a bounded group of trusted individuals or experts (Gustetic et al., 2014).

A handful of recent examples demonstrate the potential of leveraging the crowd for scientific purposes. In 2017, the "Zooniverse" project discovered a solar system

containing five exoplanets using crowd participants, referred to as “citizen scientists” (Greicius, 2018); the project also allows everyday internet users to scour and sort through high-resolution images of the lunar surface, among other complex “people-powered” projects, as a way of charting the moon’s numerous details (“Zooniverse,” 2018). Frito-Lay has recently held annual “Do Us a Flavor” contests in which the public is invited to come up with flavors for Lay’s potato chips (Burlingame, 2017). To date, research on crowdsourcing has focused largely on the utility of the collected data and its reliability vis-à-vis consistency across differing samples and types of data (Goodman, Cryder, & Cheema, 2013; Zhao & Zhu, 2014; Paolacci, Chandler, & Ipeirotis, 2010; Hsueh, Melville, & Sindhwani, 2009), and researchers have repeatedly called for further investigation into crowdsourcing’s deeper potential.

Recent studies (Mohammed & Turney, 2013; Benoit, Conway, Laver, & Mikhaylov, 2012; Snow, O’Connor, Jurafsky, & Ng, 2008; Galloway, Tudor, & Vander Haegen, 2006) have begun to assess the feasibility of crowdsourcing content analysis itself. In other words, modern research has already started to investigate the delegation of traditional roles of scholars and other SMEs as data analysts, roles often simply referenced as “coders” or “raters” in this context, to non-expert internet users. Typically, this is done in exchange for a small payment on a dedicated crowdsourcing platform (i.e., website), and results have been promising, insofar as the types of ratings requested are reasonably intuitive and briefly explained, such as the evaluation of a passage for conservative tone or for liberal tone based on opinions expressed around select keywords (Benoit et al, 2012; Snow et al., 2008). In response to the referenced calls for deeper study, the following, and subsequent line of, research seeks to extend this area of inquiry

to “fuzzier,” less popularly or commonly understood constructs (Zadeh, 1965; McCloskey & Glucksberg, 1978) which might conventionally be viewed as abstract, latent, and generally inaccessible to the public absent some level of training, rigorous study, and/or careful calibration.

CHAPTER 2: HYPOTHESIS DEVELOPMENT

2.1 A Brief History of Content Analysis

Over the course of the 17th century, numerous dissertations put forth by religious scholars investigated different means of interpreting newspaper texts. Their intent was to track the dissemination of newly mass-produced writings, and any messages the public might receive, harmful or otherwise, that were not part of the Bible (Groth, 1948; as cited in Krippendorff, 2013); their work represented the earliest recorded instances of what is commonly referred to as "content analysis" by today's empiricists. One example of how these practices gradually assumed the form of the content analyses and interpretations of qualitative data now widely in use: In the 18th century, debate erupted in Sweden surrounding a quantitative analysis of a collection of hymns. Some scholars insisted that dangerous ideas were encoded in the songs, while others dissented. Their respective sides gave rise to an academic debate over differing interpretations of qualitative data that, in many ways, continues to this day (Krippendorff, 2013).

In an effort to legitimize content analysis as an empirical technique, in the early 20th century, Eugen Löbl published a scheme of classification for the "inner structure of content"; he used newspapers' various influences upon popular culture and society in general as his primary source (Krippendorff, 2013). These processes were later given the name "content analyses" (Waples & Berelson, 1941), and their inadequacies—for instance, a dearth of theory-driven research questions in favor of more anecdotal inquiries (Krippendorff, 2013)—started to be addressed in earnest. While this was by no means the only driving factor, early psychologists and sociologists subsequently began to pursue more nuanced, rigorous work pertaining to the analysis of written communications, such

as investigations into social stereotypes and others (Lippmann, 1922; Simpson, 1934, Walworth, 1938; Martin, 1936; as cited in Krippendorff, 2013), as well as to the analysis of survey data. Content analysis began to leave the scrutiny of newspapers behind and endeavored to examine attitudes and symbols more broadly, this while operating with theoretical foundations and statistical significance emphasized for the first time (Krippendorff, 2013).

In the years following World War II, content analysis had begun to propagate into scientific disciplines beyond psychology and sociology (Krippendorff, 2013). The first scholars in communication studies founded their discipline upon these and other empirical methods of studying human communication, which helped to further elevate content analysis as not only a respected technique, but as an interdisciplinary technique. Psychologists thereafter began to test rudimentary approaches to computerizing content analysis (Stone, Bales, Namenwirth, & Ogilvie, 1962; Schank & Abelson, 1977), and from that point forward, the quantitative coding of complex textual inputs by human raters became a vital aspect of research methodology. More recently, Schnurr, Rosenberg, and Oxman (1992, 1993), Zeldow and McAdams (1993), and others (Rourke & Anderson, 2004; Morris, 1994) have performed practical comparisons between the more modern and advanced content analyses done via computers versus those done by researchers using the more time-tested manual approaches, seeking to show the fundamental comparability between the older and newer forms of content analysis.

Today's content analyses have thus benefited from years of perpetual refinement. Within the last decade, Krippendorff (2013) cited six main categories that had developed differentiating how content analysis is applied in contemporary settings: *Extrapolations*

(i.e., deducing hidden or otherwise unobserved meanings), *Standards* (i.e., making qualitative comparisons and judgments), *Indices and Symptoms* (i.e., observing and cataloging connections between phenomena), *Linguistic Re-Presentations* (i.e. discerning the means by which different meanings can arise from similar sets of words), *Conversations* (i.e., studying the complex messaging shared in speech-based interaction), and *Institutional Processes* (i.e., discerning group- and organization-level phenomena which may be less clear to individuals by virtue of levels differences). For the purposes of this paper, note that all of these categories deal with potentially abstract or implicit information that is, by its very nature, a challenge to interpret, let alone quantify.

For example, a researcher might have paragraphs of written participant responses from a study which need to be interpreted in terms of a simple numerical value. This value could be attempting to quantify a certain quality, or underlying construct, that is implicit or expected in the data (e.g., the degree of negativity in one's self-talk). Or perhaps a small team of researchers might need to count the number of times a specific mode of verbal or written expression, such as sarcasm, is used in a body of recorded speech data containing hundreds of hours of content for the researchers to pour over. In this case, the researchers on the team would have to decide in advance, or upon preliminary review, what types of words or expressions indicate sarcasm in the study's particular context.

In the event of a large dataset, such preparation and evaluation of data could take considerable time, and the above is by no means an atypical example. While today's content analysis is more efficient than it was in the 1980s and before (Krippendorff, 2013), an interdisciplinary array of researchers have continued to acknowledge the time-

consuming side of the process in papers published as recently as 2014 (Cho & Lee, 2014; Gale, Heath, Cameron, Rashid, & Redwood, 2013; Hopkins & King, 2010; Pope & Ziebland, 2000 Rourke, Anderson, Garrison, & Archer, 2007; Wilkinson, Joffe & Yardley, 2004). While some of these authors cite different aspects of the process as the most time-consuming relative to others, one general consensus between them is that content analysis takes an amount of effort and time often disproportionate not only to the rest of the project, but to other modes of research in general. Thus, it is helpful to examine, in an ongoing sense, innovative approaches such as crowdsourcing which may speed the process, as well as reduce a common and time-consuming burden on researchers, without sacrificing validity.

2.2 The Crowdsourcing Surge

In response to these and other, similar challenges, "the crowd" is becoming an increasingly popular solution for the demands of modern data collection (Bohannon, 2011). Prior to showing the connection between the crowd and content analysis, some fundamentals surrounding the concept of the crowd and crowdsourcing as a viable data collection technique should be addressed. Jeff Howe's (2008) original definitions of the crowd and crowdsourcing remain widely cited in the literature: The crowd is defined simply as a large group of people requiring no qualifications for group membership. Crowdsourcing, by extension, is viewed as providing the ability to "facilitate the connectivity and collaboration of people, organizations, and societies" to collect and analyze data (Zhao & Zhu, 2014, p. 1). While there is still not a very deep literature at present, and while there is still variation on terminology across journals, support has been

demonstrated for the validity and reliability of the data collected via the crowd in general, given certain conditions.

For instance, consider the initial steps an author of a study using crowdsourced data must take. They must demonstrate, with some authority, that any data collected from the crowd can reasonably be expected to compare to data collected in the traditional domain of their area of expertise. While such comparisons are not always plausible and thus not all crowdsourced data is created equal, several papers have reported positively on the possibility of such fundamental similarities between crowdsourced and traditionally collected data (Paolacci & Chandler, 2014; Paolacci et al., 2010). In some cases, usage of the crowd in the place of traditional sources even seems to improve data quality; for example, Paolacci et al. (2010) showed that, when conducting research online, participants recruited through Mechanical Turk in particular were more likely to see a survey task through to completion than other internet-based participants.

Others have delved more deeply into the effects of soliciting data from an open platform, such as the accuracy of reported demographic information. David Rand (2012) found that, when testing the accuracy of the crowd's self-reported demographic information, 97% of responses were correct. Snow and colleagues (2008) compared crowdsourced content ratings to traditional expert ratings by comparing the mean ratings of each group; also conducted were more robust analyses comparing how many non-expert ratings (i.e., crowd ratings) were required to compare to those of experts. They found high similarity for ratings in the categories of affect recognition, word similarity, recognizing textual entailment (the direction of the relationship between fragments), event temporal ordering, and word sense disambiguation (the process of clarifying how

certain ambiguous words are specifically meant in particular usages). Horton, Rand, and Zeckhauser (2011) pointed optimistically to the ability of crowdsourced samples to replicate traditional, experimental scientific findings in general, focusing on laboratory-style crowd study and finding evidence for both internal and external validity.

Despite the crowd's many successes, Goodman et al. (2013) point out that crowdsourced study participants' attention spans are generally lower and thus more controls (or screens) are needed to ensure data quality. For example, Hsueh et al. (2009) provide three criteria for reducing data problems associated with unreliability: (1) pre-empting noise in collected data by incorporating means to improve the reliability of responses into study design, (2) assessing the degree to which content being analyzed by the crowd is difficult for participants to rate or to categorize definitively and making accommodations where possible, and (3) lexical uncertainty, or the presence of poorly structured or poorly worded content which might directly lead to such difficulties as in (2). Furthermore, in the 2008 study by Snow et al., the authors highlighted the importance of ensuring reliable crowds by making payment contingent upon quality responses. They further recommended using site-specific controls designed to help screen participants based on the acceptance / rejection rates of past responses. These findings reinforce a growing consensus that crowdsourcing can be both effective and reliable, but only when done well.

2.3 Crowdsourcing Content Analysis

Just as crowdsourcing has been shown to be useful as a tool for data collection, in that crowdsourced projects can produce similar results to projects employing traditionally-collected data, recent studies in more than one field have begun to support

the proposition that crowdsourcing can likewise be useful in the replication of traditionally-conducted content analyses. This is to say that the crowd, given the aforementioned quality controls, can not only provide similarly useful data to traditional survey participants, but can also analyze data similarly to traditional means. When Hsueh et al. presented annotation data on clips from political blogs to a team of expert raters and a team of non-expert raters, there was approximately 84% agreement between the groups (2009) (though their method for determining agreement was not given). In other words, there is potential for the crowd to not only take the place of some traditional research participants, but some traditional analytical tasks as well.

Moreover, a high ratio of non-expert to expert raters may not be required to achieve high reliability, contrary to what might be assumed. Benoit and colleagues argued, quite straightforwardly, that a group of experts (such as graduate or undergraduate research assistants making assessments of data) is no less a crowd than a larger group of non-experts in many common research contexts. To achieve a valid proxy to the judgment of actual subject matter experts and other "traditional" raters, all that may therefore be needed is to have a moderately-sized sample of non-expert ratings (2012) and a dataset that can be considered reasonably intuitive by non-experts in the subject area. For example, Snow and colleagues ran group-by-group comparisons of non-expert ratings to expert ratings, steadily incorporating more non-expert ratings into the non-expert mean while repeatedly comparing, and found that, contrary to expectations, four non-experts could, on average, rival the judgment of a single expert in natural language annotation tasks (2008). Relatively unstudied, however, are the qualities in a crowd which might make some crowds behave and rate more similarly to expert raters.

As such, given the number of coders readily accessible via crowdsourcing, and provided that the task is accessible, provides clear instructions, and employs empirically validated quality controls, the crowd might be able to collectively process even more complicated information and solve even more complex problems than we have yet imagined. Further justification for pursuing this research question stems, in part, from the laws of statistics: As any given sample of crowd participants grows, sampling error is reduced and the average statistic generated by the sample will approach the greater population's average rating parameter. If there exists a reasonable expectation that the greater population can collectively understand complex data, the most sensible and straightforward step to take next is to determine how many non-expert participants are required at any given level of task accessibility, clarity, quality control, and so forth.

Consider this fifty-year-old example: After the 1968 disappearance of the U.S. submarine *Scorpion*, a group of non-experts was tasked with combining their diverse knowledge to assist in finding the ship. The search area had been deemed hopelessly broad, over twenty miles wide and many thousands of feet deep (Surowiecki, 2004); undeterred, a naval officer concocted a strategy that involved pooling a team of men with varying backgrounds—from mathematicians to salvage workers—and asking them to each come up with an estimation of where the ship might be, without collaborating with one another. After using Bayes's theorem to combine their guesses into a composite estimate, the submarine's final resting place was found just 220 yards from the predicted spot (Surowiecki, 2004). Given these many examples, and given the gradual minimization of error in increasingly larger crowds expected under Classical Test Theory

(Novick, 1966; Lord, Novick, & Birnbaum, 1968; Allan & Yen, 2002), it is arguable that the crowd may be able to tackle more complex content analysis and ratings tasks.

2.4 The Question of Expertise

Unsurprisingly, the world of data science has been experiencing a wave of crowdsourcing mechanisms and websites in recent years. One of the most well-known crowdsourcing mediums is Amazon Mechanical Turk, frequently shortened to M-Turk. Designed as "an interface for creating jobs, combined with a massive global pool of workers" (Benoit et al., 2012, p. 11), M-Turk provides online users with a small amount of compensation provided by the requester (the site's term for the person or organization running the survey) in return for taking the requester's survey, or Human Intelligence Task (HIT or task for short). The amount of compensation is usually contingent upon the amount of work, and has ranged from less than 50 cents to amounts rivaling the equivalent of U.S. Federal Minimum Wage for the total time one is expected to take to complete the task. As mentioned previously, recent studies have evidenced the usefulness of crowdsourcing platforms to the social and organizational sciences (Paolacci, Chandler, & Ipeirotis, 2010) as a means of data collection; Some utilized M-Turk, while others used a platform similar to it, such as CrowdFlower (Benoit et al., 2012; Finin et al., 2010, Zhai et al., 2013).

In spite of the mounting evidence that deep subject knowledge and traditional expertise might not be essential, in a blanket sense, to the time-intensive content analysis projects which typically call on such sources, how we qualify and filter crowd users for participation in studies is nevertheless of vital importance. A common assumption made by all of the studies mentioned thus far is that the non-experts in those studies possess

some level of capability with the task expected of them; despite a lack of traditional expertise, they are capable of performing an acceptable proportion of a task, or performing a task to an acceptable degree of accuracy. Imagine a research project in which non-expert, crowdsourced participants are provided with a classification guide for a ratings task. The guide helps them categorize video content based on emotional tone. Accordingly, the guide might contain key visual or auditory (assuming sound is used) indicators which help participants decide that a video belongs in a certain emotional category, such as ecstatic, morose, affectionate, or resentful.

The studies cited above support the notion that, with a relatively minor degree of preparation, the crowdsourced participants would be able to successfully categorize the content presented to them in such a research project. While the example makes no assumption that the crowd possesses traditional expertise on emotional interpretation, it employs the more justifiable assumption that the crowd is capable of using past experiences with video content (TV, YouTube, etc.) as a valid form of expertise with which to complete a research task. In the same way, crowdsourced projects which feature traditional opinion surveys or similar questionnaires have a common tendency to assume that their participants have answered electronic questionnaires on prior occasions.

Furthermore, when considering this question of expertise, the mutual control both researcher and participant have over crowd participation through crowdsourcing platforms like M-Turk should not be overlooked (Snow et al., 2008). For some time, tools to help researchers obtain results from “veteran” raters have been made available on crowdsourcing platforms. M-Turk calls raters in these categories “Master” raters, a distinction obtained by high performance in “statistical models that analyze Worker

performance based on several Requester-provided and marketplace data points" ("Amazon Mechanical Turk," 2018). Additionally, respondents themselves have the freedom to choose surveys they feel comfortable taking, and avoid those they might not understand or have time for. Crowdsourcing platforms normatively allow participants to choose tasks based on the type of work involved, and after reading a summary provided by the creator of the questionnaire; these controls increase the likelihood of receiving participant data from both willing participants and participants who feel comfortable with the subject matter. The improper usage or implementation of these controls can open a proverbial floodgate to the kinds of situations which have given so many researchers pause when considering crowdsourced data collection or data analysis (Cheung, Burns, Sinclair, & Sliter, 2016; Kittur, Chi, & Suh, 2008).

One reason for this is that "spammers" and other crowd-based interferences, such as cultural variations, can create noise in a dataset, large amounts of which render data useless. The present incarnation of Amazon M-Turk and CrowdFlower have robust filtering options compared to the available tools of 2008 and even to just a few years ago (Benoit et al., 2012). Called "qualifications," M-Turk allows user-created, user-managed requirements for performing crowdsourced tasks ("Creating and Managing Qualifications," 2017). Consequently, researchers employing M-Turk to collect or analyze data could not only require that an M-Turk worker trying to access a survey possess certain professional experience or be of a certain age, but require M-Turk to filter out individuals who have not demonstrated reliable responses after performing other surveys and tasks on the website. What follows is to discover more concrete evidence for

what types of content are best to analyze via the crowd, and which qualifications and prerequisites are most useful for honing crowdsourced data quality overall.

2.5 Rating Fuzzier Constructs

Though extant research shows crowds to be capable of making comparable judgments to those of experts on tasks such as discerning the political orientation of a text (Hsueh et al., 2009), and even on less intuitive natural language tasks such as affect recognition, word similarity, or the temporal ordering of events (Snow et al., 2008), a wide variety of questions remains to be answered. This is especially true for constructs that are more traditionally left to those specializing in the social sciences. One category of data in this domain is the category of "fuzzy" constructs. These are constructs without logical boundaries and which are frequently assumed to require some degree of education in psychological, sociological, communication, or information science in order to be understood and assessed accurately. Fuzzy constructs are rarely part of popular discourse or even considered public knowledge.

"Fuzziness" is a quality held by a wide variety of data both outside and within the social sciences. Coined in the Information Sciences literature by Zadeh in 1965, the term refers to any set of data for which the logical boundaries cannot be clearly defined. In psychology as well as statistics, the related term "latent" commonly applies to variables for which a value is implicit, approximated, or below the surface, rather than explicit, precisely measurable, or surface-level. As the majority of constructs measured in psychology cannot be directly observed, they are often categorized as "latent" variables or constructs in that they are approximations of precise values; as they likewise lack logical constraints, they can also be considered "fuzzy." Such "fuzzy" variables and

constructs generally require deeper explanation; for example, how to accurately and fairly assess the extent to which someone is open to experience, conscientious, extraverted, etc.

Facilitating—and, indeed, receiving—training on rating such fuzzy data generates considerable work hours for researchers and graduate students, in general because these datasets and constructs are taken not to be interpretable by those without training or experience, or who are not subject matter experts (SMEs); thus, the option of outsourcing the work to non-experts is seldom discussed. Content analysis typically involves extensive reliability training, or calibration (Krippendorff, 2013; Hsieh, 2005; Kolbe & Burnett, 1991). The process of training raters in this way is especially time-intensive when dealing with logically unconstrained constructs, and the coding itself is often performed by graduate students in such instances, or even by principal researchers themselves (Krippendorff, 2013). The question of whether fuzzy data, with "fuzzy" representing a level of above-average complexity that might still be to some degree approachable by non-experts, can instead be rated by the crowd instead would be a considerable advancement in crowdsourcing research.

2.6 The Present Study

This project therefore sought to explore whether the crowd can reliably evaluate constructs which most often find themselves assigned to academics, students, researchers, practitioners, or other SMEs, and to delve deeper by assessing how accurately they are able to rate such constructs when compared to such SMEs. Analyses, particularly of accuracy, were facilitated by employing an archival dataset created around a fuzzy construct. The construct chosen was, on the surface, "fuzzier" in comparison to constructs in recent crowdsourced content analysis studies, such as political orientation or

word similarity in a text. Crowd ratings of the construct were then compared to SME-generated ratings of the data from the dataset's associated study. By sampling a much larger crowd than the SME group from the extant study, the crowdsourced sample could be partitioned into different sizes during analysis, providing useful perspective on how crowdsourced ratings' accuracy changes between different crowd sizes, and which group—SMEs or the crowd—is more accurate or reliable at various sizes. Exploratory analyses on individual crowd participant data sought insights on which coder qualifications appeared to drive reliability and accuracy over and above typical screening methods (e.g., filters such as site-reported coder quality, hurried responses, age, locale, etc.).

Constructive self-talk, more specifically the constructiveness of a passage of self-talk data from the extant dataset, was chosen as the fuzzy construct. Constructive self-talk is defined as "[conveying] a rational and nuanced understanding of oneself or a situation; [viewing] obstacles in the environments as challenges, as opposed to threats; generally [including] motivational and/or instructional language; […] usually optimistic, without being naively so" (Uhrich et al., 2017). This is a quintessentially fuzzy construct in that "hard" logical boundaries of self-talk qualities such as constructiveness are indiscernible. As concepts, both constructiveness and self-talk lack boundaries which are logical, or concrete in some discernible, explicable way. While it is possible for different raters to evaluate them similarly through calibration and practice, it is generally not possible to completely eliminate minor subjective differences between raters. This is especially the case in larger samples of ratings and on ratings of constructs such as constructive self-talk, whereas—for example—the simpler, more concrete task of training two raters to

identically assess passages by counting pre-defined keywords can be accomplished with near-perfect efficacy. When seeking to determine whether a new group of raters is reliably rating constructive self-talk, it is critical to compare ratings from the group to an experienced group's ratings in terms of reliability. To also compare the two sets in terms of accuracy, however, would require at least one additional variable with which to support the assertion that either set of ratings is accurate.

The distinction between accuracy and reliability is particularly important for the purposes of this study. Only after reliability has been established in a group of either expert or non-expert raters should the subsequent assessments of accuracy be considered meaningful. While reliability statistics demonstrate that groups of raters perform similarly across time or between different groups, assessments regarding accuracy are not always possible if there is not any additional means to infer accuracy from the data. When such assessments are possible, however, they can exceed the utility of simple reliability comparisons by demonstrating the likelihood that one set of ratings is close to the true value of the construct in question than another. It was therefore crucial that an appropriate means for inferring accuracy—in this case, self-reports of perceived self-efficacy associated with the passages of self-talk in the archival data—be present in the dataset.

According to Social Cognitive Theory (Bandura, 1986), one's self-assessment of their ability can have significant effects on performance. His more recent research draws an even more explicit connection between self-management processes, such as self-talk, and self-efficacy (Bandura, 2001): Whenever failures are treated as challenges to be overcome, a key component of constructive self-talk as defined above, it tends to lead to

the redoubling of efforts as opposed to diminishing efforts (Bandura, 2001). According to Bandura, one's beliefs surrounding one's efficacy can be shown to predict adaptive responses to one's environment (such as the content of self-talk), but also go farther by directly influencing the levels of optimism and pessimism in one's thoughts (2001). Efficacy beliefs lead us to select which "challenges to undertake, how much effort to expend in the endeavor, how long to persevere in the face of obstacles and failures, and whether failures are motivating or demoralizing" (Bandura 2001, p. 10). It was therefore expected that constructive self-talk ratings from crowdsourced participants would correlate significantly with self-reports of perceived self-efficacy among the self-talk participants in the archival dataset, and that the strength of this relationship could be interpreted as support for the accuracy of the crowdsourced ratings.

There are a number of reasons, some theoretical and some based on recent empirical studies, that we expect crowds of certain sizes might be capable of producing comparably accurate judgments of such data to SMEs. The previous sections cover a variety of recent studies that have successfully demonstrated that non-traditional crowds can effectively rate several types of more concrete data traditionally rated by SMEs (Benoit et al., 2012; Galloway et al., 2006; Mohammed & Turney, 2013; Snow et al., 2008), given certain controls (Hsueh et al., 2009), and that such crowds need not be much larger than typical groups of SMEs rating the same data (Benoit et al., 2012; Snow et al., 2008). Given the nascency of the subject matter, there is no independent and strongly supported theory of crowdsourcing; however, the Central Limit Theorem states that distributions of sample means from a larger population (of raters, in this case) become increasingly normally distributed as sample size approaches infinity. It is therefore within

reason to assert there is a size of a group of crowdsourced raters where statistical comparability to SME raters of similar content—though crowd groups are likely larger on average than SME groups—becomes possible.

Our inquiry followed, then, as to what the smallest possible crowd might be that can accurately code fuzzy data (Greengard, 2011; Bonabeau, 2009), with one caveat: the variable of interest must be one that the crowd could be expected to code with an acceptable level of fidelity.

RQ1: At what size, if at all, can groups of crowdsourced raters of a "fuzzier" construct approximate the reliability and accuracy of ratings generated by Subject Matter Experts?

Furthermore, implied in the comparison of a non-expert group to an expert group is the question of specific qualities which allow non-expert ratings to achieve comparability to expert ratings. The expert raters participating in the archival study were doctoral-level students with an inherent desire to make quality contributions to the project. Assuming the ratings of specific group sizes from this crowd sample compare to the accuracy of traditional/SME ratings, filtering the crowd based on certain individual differences—among those reasonably expected to have affected reliability and accuracy on the ratings task for the SMEs, such as motivation, engagement, level of education, and previous experience with similar tasks—might therefore improve upon the crowdsourced raters' comparability to traditional/SME ratings.

RQ2: Can rater motivation, rater engagement, level of rater education, and previous rater experience be used as filters in order to improve crowd accuracy and reliability?

CHAPTER 3: METHOD

3.1 Sample

Analyses drew from three sources of data:

Data source 1. Qualitative self-talk data (in the form of passages) and quantitative self-report rating of academic self-efficacy (to be used for accuracy assessment). Provided by 177 archival participants in a prior self-talk study.

Data source 2. Archival SME ratings of the self-talk data provided by the archival participants mentioned in data source 1. There were a total of six SMEs in this ratings group.

Data source 3. Crowdsourced ratings, collected from Amazon Mechanical Turk (M-Turk), of the qualitative self-talk data from data source one. All ratings were collected from verified M-Turk "Masters," the platform's current label for site-verified high-reliability respondents (n = 96), and the ratings tasks were only made available to participants in the United States. Participants on M-Turk are required to be 18 years old or older in order to set up an account.

3.2 Measures

Constructive self-talk. The self-talk data provided to all participants for ratings came from a prior study—data source 1—which used a research pool of undergraduate psychology students at a small private college in the southeast (n = 177). Participants provided samples of self-talk in response to a set of imagined scenarios (for clarity, "archival participants" or "archival raters" will henceforth refer to the aforementioned student participants from data source 1, and "crowd participants" or "crowd raters" will refer to crowdsourced participants from newly collected data source 3). These scenarios

were designed to examine the archival participants' responses to stressful situations by soliciting possible self-talk responses to that scenario, for example:

> *Think about an academic challenge that you are currently experiencing (e.g., a difficult class, a hard assignment, etc.). Stop reading and focus on the kinds of thoughts that go through your head when dealing with this challenge for 30 seconds. In a sentence or two, briefly describe the challenge. Next, please write down the unedited dialogue that runs through your mind (i.e., thoughts) when you are thinking about this challenge. Be sure to write in the first person, "I am thinking…" Please write at least a few sentences.* (Uhrich et al., 2017)

This scenario yielded a variety of responses from the archival participants. The following is a sample response from an archival participant.

> *"I have an essay for my German film class due in a week in which I have to analyze a film and support my argument with sources. I am thinking that I really want to do well because the last class that I had with this professor was taught in German, at too high of a level, and I gave up at the end of the semester and disappointed myself and probably my professor in how bad my performance was at the end of the semester. I know I can do it; it's just a matter of sorting out the cloud of thoughts and take the time to write a coherent essay."* (Uhrich et al., 2017).

This particular scenario's SME ratings (from data source 2) had the highest r_{wg} ratings—0.89—of inter-rater reliability of any of the varying scenarios presented in the archival study. As this set of ratings was the most reliable of the different scenarios rated by the SMEs, passages from only this scenario were presented to crowd participants in this

study. Crowd raters were broken into groups (of 12), each group rating evenly-divided sets of passages (20). This was a similar method of assignment to the "batches" of work originally assigned to the SME raters, both done with the intention of distributing the work among the group (n = 96; 8 groups of 12 coders rating 20 passages out of a pool of 160). This did place an upper limit on the number of passages used, as using all 177 would have resulted in an uneven distribution of work for the final ratings group, (the first 160 were therefore used instead).

The original SME ratings of the self-talk samples' constructiveness were performed by a group of doctoral-level raters (n = 6) trained in qualitative methods as well as the self-talk content area. Self-talk of the archival participants from data source 1 was assessed for constructiveness. After providing a detailed definition of the constructive self-talk dimension, raters provided a single rating on a 5-point Likert scale ("1 - No evidence of self-talk dimension"; "2 – Little evidence of self-talk dimension"; "3 – Some evidence of self-talk dimension"; "4 – Good evidence of self-talk dimension"; "5 – Great evidence of self-talk dimension"; Uhrich et al., 2017). SMEs were instructed to read and assign a rating to participants' entire responses instead of scrutinizing passages by sentence. This was important for all raters in both the archival and present study, as the overall, or average, contextual nature of the entire passage was the primary coding unit, rather than specific sentences, and furthermore because constructiveness sometimes varies from one sentence to the next (Uhrich et al., 2017), ergo a single rating for multiple sentences can be more reliable than multiple ratings per passage. The independent ratings from each SME were then aggregated to create a mean constructiveness score for each scenario and each participant.

Academic self-efficacy (ASE). In order to demonstrate the validity of ratings performed on their self-talk, archival participants also reported their level of academic self-efficacy at the time of the survey. A seven-item scale (Greene, Miller, Crowson, Duke, & Akey, 2004) was used to assess academic self-efficacy (ASE) ($\alpha = .94$) (See Appendix B). The original scale was designed to rate self-efficacy in a single class, whereas the prior study changed the items to rate self-efficacy across all classes, thus capturing a more general sense of overall self-efficacy in academic/collegiate settings. Items were placed on a 5-point Likert scale (from "1 – Strongly disagree" to "5 – Strongly agree").

As in the archival study, correlations between this scale and SME and crowd ratings for constructive self-talk will be used as an indicator of accuracy, respectively, due to the strong theoretical linkage between these two constructs. There are a number of interconnected explanations one might invoke to support the usage of ASE as an indicator. Beginning with face value, constructiveness and ASE are theoretically related concepts. Constructive self-talk, by the definition supplied, is arguably a verbal articulation of one's perceived self-efficacy, or is close to being so (low self-efficacy is likely to be associated with self-talk low in constructiveness while high self-efficacy is likely to be associated with self-talk high in constructiveness). While the constructs are not interchangeable, strong and significant correlations between them (where ASE is the dependent) can be interpreted as support for the similar accuracy of the constructive self-talk ratings between SMEs and the crowd. A stronger correlation for one group (SME or crowd), given strong reliability statistics, could even indicate higher accuracy in one group.

Furthermore, the archival participants in our study (data source 1) are all students, thus are all more likely to provide accurate self-report of their academic self-efficacy than non-students, given the salience and intuitiveness of the scale; similarly, the responses from the chosen scenario prompt strongly invokes the responder's sense of ASE in that it prompts the archival respondent to imagine an end-of-semester studying crisis; accordingly, SME or Crowd ratings of the constructiveness of a passage of self-talk written in response to a prompt about such a crisis can be expected to relate significantly to self-report ratings from the same student about their academic self-efficacy. Finally, a significant correlation exists in the archival study between the SME constructiveness ratings and the archival participants' ASE self-reports, therefore the presence of such a correlation between crowd ratings and ASE self-reports is an additional indicator of data similarity, over and above reliability assessments, by which accuracy can be inferred.

Intrinsic motivation. Motivation was assessed in crowd participants from data source three using the 18-item Intrinsic Motivation Inventory (IMI; $\alpha = .78$) (McAuley, Duncan, & Tammen, 1989; Ryan, 1982; for list of items, see Appendix C). McAuley and colleagues condensed the original 27-item IMI—which was designed to be shortened as necessary—to 18 items. These items assessed four dimensions contributing to intrinsic motivation (interest-enjoyment, perceived competence, effort-importance, and tension-pressure). A minimum of four items were maintained per subscale in the 18-item scale's design. All items were reworded from the 1989 version, which focused on basketball, to focus responses upon a ratings task. All items were included to create overall motivation scores (with tension-pressure reverse scored as appropriate).

Task engagement. Task engagement was measured by the 7-item task engagement portion of the Short Dundee Stress State Questionnaire (DSSQ; Matthews, Emo, & Funke, 2005). The original DSSQ was created using an eleven-factor model, with each of the eleven stemming from three overarching constructs: task engagement, distress, and worry. Four subscales—energetic arousal, task interest, success motivation, and concentration—contributed to overall task engagement assessment. Cronbach's alphas were .80, .75, .87, and .85, respectively. The shortened DSSQ provides a more expedient alternative to the lengthy DSSQ, and the Engagement portion of the questionnaire can be presented independently, as it was for the purposes of this study.

Levels of education & experience. Participants were asked to report their level of education ("Select the response which best describes your level of education") on a 5-point, one-item scale ("1 – No high school diploma or degree"; "2 – High school diploma"; "3 – Some college, associate degree"; "4 – College degree, B.A. or B.S."; "5 – Advanced or professional degree"), and to indicate the extent to which they possessed applicable past experience in similar exercises. The question read, "To what extent do you feel you already have experience with similar tasks to the ratings task you performed for this survey?," (from "1 – Lowest" to "5 – Highest").

3.3 Data Collection

The crowdsourced sample (source 3) was collected in waves using Mechanical Turk. Eight surveys (HITs/tasks on M-Turk) were administered, each identical in design but containing 20 completely different self-talk passages from data source 1. Two additional passages randomly chosen from the archival data's last 17 passages, otherwise

unused in the surveys, were offered across all 20 surveys for exploratory analyses. Duplicate participation was monitored and rejected.

Consistent with past studies on ensuring validity through crowdsourced ratings (Benoit et al, 2012), appropriate controls (e.g., restrictions on the participation of unreliable participants as assessed by past participation on M-Turk) were applied to ensure the reliability of participants. As crowd participants will perform ratings on the existing self-talk scenario data in the same way the SME group already has for the archival data, SMEs from the archival study were consulted on the average length of time expected for the size of a 20-rating task (15 minutes). Accordingly, any participants who submitted surveys significantly faster (less than 10 minutes) were rejected and new participants were sought until 12 participants have taken each questionnaire. Participants who were determined to have participated in other iterations of the same survey were likewise rejected. A total of six participants altogether were rejected on these grounds.

Participants in each group were not aware of their group (or HIT/task) number and only saw the relevant set of passages given to their group. Justification for a group size of 12, as opposed to more or fewer, was as follows: The original Uhrich et al. study (2017) used a team of six SMEs, all of whom rated self-talk passages in different combinations of three team members (i.e., no more than three SMEs rated any one self-talk passage). In order to test reliability and accuracy with different numbers of crowd raters on the same passages, desired team size (6) was doubled for the crowd.

As a means of assessing the crowdsourced sample for the presence of individual-level variations which might have contributed to the accuracy of their codes, crowd participants then completed the motivation, engagement, education, and experience

questionnaires. Participants responded to these items after their self-talk ratings were complete, but before exiting the task.

3.4 Procedure

Randomly selecting crowds. Analyses first assessed groups of crowd raters at a size of 12 raters; these analyses included all 96 crowd participants. Analyses then assessed incrementally smaller groups to ascertain the group size at which reliability emerged. Groups were analyzed at 9, 6, and 3 raters, each iteration of the analysis using random selection to remove participants. Smaller groups were assessed only after each larger group demonstrated reliability and/or accuracy in their responses. If the last group assessed was not fundamentally reliable by more than one measure, analyses were not pursued for smaller groups. Note that the groups are intentionally referenced as "raters groups" as opposed to "rater groups." As eight separate groups of 12, 9, 6, or 3 raters were merged to assess crowd ratings for reliability and accuracy, this distinction can be helpful. For example, a total of 48 raters, eight groups of six, were involved in the "six-raters" analyses.

Crowd raters first and foremost needed to demonstrate reliability. Within-group consistency was assessed using Cronbach's alpha and intra-class correlation coefficients. These tests required ratings from "Rater 1," "Rater 2," "Rater 3," etc., to be merged across all eight groups of raters (in each of which there was a different Rater 1, Rater 2, Rater 3, etc.) in order to assess the reliability of the entire sample. For example, the original eight groups of twelve raters that participated in the present study each had six members selected at random to create eight groups of six (as mentioned above). The first selected rater in the first group shared the same row with the first selected rater in the

next group, and so on until every rating was accounted for and the entire crowd's work could be assessed using a predetermined number of raters. The same type of merging was performed in order to generate means for each passage of self-talk rated by select numbers of crowd raters. This might provide helpful context into how the "six-raters" group actually constitutes 48 individual raters contributing to six lines of rater data. For a visual of how all eight groups of twelve raters made up the "twelve-raters" group, see Table 2. Note that Table 2 abbreviates each set of 20 passage ratings, the full list of raters participating, and the number of tasks shown for simplicity and clarity.

The archival study from which the SME data was drawn made use of a similar approach by using mean data from three SMEs (out of six total); various combinations of three out of the six SMEs were assigned to rate all of the different passages of self-talk. As a result, each mean rating for a given section of the data (such as the 20 ratings assigned to each crowd group in the present study) was based on ratings from a group of SMEs. For the purposes of statistical analyses, the data therefore considered the SMEs interchangeable with one another, proceeding from the assumption that the mean of any three SMEs' ratings would be as useful as the mean of another combination of SMEs. The present study, therefore, treated its participants in the same way by using the described methodology. Merging rater data across groups in this way was also necessary because an individual group's task size (20 ratings) might be too small to achieve the desired level of reliability; this was confirmed in exploratory analyses of both SME and crowd rater groups.

Chi-square tests were conducted to confirm statistical similarity between coefficient alphas (Diedenhofen & Musch, 2015). Given the random nature of

participants who volunteered to supply crowd ratings, ICC estimates and 90% confidence intervals were calculated based on a consistency-based, one-way random model. Between-group consistency (consistency between the crowd raters and the SME raters) was measured by correlating the constructiveness scores of the two groups, and inter-rater agreement was assessed for each group using r_{wg}. Crowd raters furthermore needed to demonstrate a similar distribution of ratings when compared to SME raters, so skewness and kurtosis data were compared.

Having successfully demonstrated reliability and similarly distributed ratings, crowd raters next needed to show they were accurate. Accuracy was assessed by correlating a raters group's constructiveness scores with the corresponding academic self-efficacy (ASE) scores from the passages rated. Fisher's r-to-z transformation was used to assess the difference between the crowd and SME accuracy correlations. If a similar level of accuracy is indicated (for instance, by a non-significant comparison between the correlations), the crowd raters groups' means then needed to be directly compared to the SMEs' means as a final indication of any similarity or difference between the two groups' ratings. A paired-samples t-test assessed the mean difference between the twelve-raters groups constructiveness scores and SME raters groups. Results from these t-tests were considered meaningful only if reliability and distribution statistics were similar.

Following successful reliability tests on crowd groups of a certain size (12, 9, or 6), the number of total crowd raters participating was reduced and the analyses repeated. To ensure smaller groups of crowd raters were first assessed with the minimum possible care taken to shape the quality of each group of raters, three crowd participants were randomly selected out of each group of raters each time until the same tests could be

performed on a nine-raters group, a six-raters group, or a three-raters group. Accordingly, each set of similar reliability and accuracy results will be increasingly meaningful the smaller the size of the crowd raters group becomes.

Filtering using performance-predicting variables. Assuming a group of a certain size smaller than twelve demonstrated reliability and accuracy across all indicators, particularly given that this would be achieved using completely random selection methods, the filtering analyses were performed only on the smallest group that yielded such results (and any groups smaller than those) and not on the larger groups. Analyzing the larger crowd raters groups when smaller groups had already shown similarity to SME ratings was deemed superfluous as the best approach to the present research questions was to identify the smallest possible reliable group.

Four distinct groups of raters were created (top-ranked raters in Motivation, Engagement, Level of Education, and Level of Experience, respectively) using participants from each of the eight MTurk HITs/Tasks. These participants made up the "Motivation Selection," "Engagement Selection," "Experience Selection," and "Education Selection" groups, using the smallest group size indicated as reliable and accurate while using random selection. Reliability / accuracy in these "criterion selection" groups were then assessed in the same way as with the analyses performed on the randomly selected groups. As an alternative to the selection-based method, the eight least-motivated of the 96 crowd raters, one from each MTurk Task/HIT, were dropped from the smallest randomly-selected group that demonstrated both reliability and accuracy. Reliability / accuracy analyses were repeated on the resulting five-raters "Motivation Elimination" group. The second least-motivated rater was then dropped and

the resulting four-raters “Motivation Elimination” group was tested. Subsequently, the same one- and two-rater eliminations and tests were conducted for one five-raters and one four-raters “Engagement Elimination” group.

For an illustration of how specific members from each of the eight MTurk task/HIT groups were chosen for these different groups during analyses, see Table 3. Bear in mind (a) that the numbers and individual criterion ratings shown are for demonstration purposes and are not real data, and (b) that analyses were performed on all eight MTurk task/HIT groups at once after each of the eight groups of raters had been selected as shown.

CHAPTER 4: RESULTS

4.1 Research Question 1

RQ1 was interested in the size of a group of crowd raters needed to achieve similar reliability and accuracy to ratings of fuzzy constructs conducted by SMEs. Results are presented in Tables 4a and 4b, with confidence intervals included for comparison. Importantly, statistically significant differences between the crowd and the SME group were rare across all analyses when it came to reliability and accuracy. Accordingly, all subsequent discussion of the reported similarities between SME and rater reliability and accuracy will emphasize two additional indicators of similarity in the quality of ratings. These are (a) whether the ratings' reliability statistics fall within generally acceptable ranges, e.g. $\alpha > .7$, ICC(1) > .5, etc., and (b) any other consistent differences between said statistics (given that statistical strength and similarity are first accounted for).

Twelve raters. Both reliability checks demonstrated consistency within groups of crowd raters ($\alpha = .80$, ICC(1) = .53) as compared to consistency within groups of SME raters ($\alpha = .92$, ICC(1) = .81). There were no sizable apparent differences between crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, $n_1 = 3$, $n_2 = 12$) = .56, $p > .05$. Ratings from the twelve-raters and SME raters groups were highly correlated, indicating good consistency between groups ($r = .76$, $p < .01$); the twelve-raters group showed good interrater agreement ($r_{wg} = .65$) as compared to SME interrater agreement ($r_{wg} = .79$). r_{wg} analyses assumed normal distribution, which was consistent with the archival study's methodology. Both distributions skewed left, with a crowd skewness of -.27 (SE = .19)

and an SME skewness of -.19 (SE = .19) accompanied by crowd kurtosis of -.97 (SE = .38) and SME kurtosis of -.35 (SE = .38). The twelve-raters group's ratings were significantly correlated with Academic Self-Efficacy (ASE; $r = .29$, $p < .01$), as were the SMEs ($r = .27$, $p < .01$). The difference between the twelve-raters group's accuracy and the SMEs' accuracy was not significant ($p > .05$). A paired-samples t-test assessed the mean difference between the twelve-raters groups' constructiveness scores (M=3.11, SD = .98) and SME raters groups (M=2.77, SD = .82). There was a significant difference between the two groups; $t(158) = 6.77$, $p < .001$), indicating significantly higher ratings among the crowd raters versus the SMEs.

Nine raters. Reliability tests found strong evidence of consistency within groups of crowd raters ($\alpha = .92$, ICC(1) = .82) when compared to consistency within groups of SME raters ($\alpha = .92$, ICC(1) = .81). There were no sizable apparent differences between crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, $n_1 = 3$, $n_2 = 9$) = .000, $p > .05$. Ratings from the nine-raters and SME raters groups were highly correlated, demonstrating good between-group consistency ($r = .76$, $p < .01$); the nine-raters group showed good interrater agreement ($r_{wg} = .63$) as compared to SME interrater agreement ($r_{wg} = .79$). r_{wg} analyses assumed normal distribution, which was consistent with the archival study's methodology. Both distributions skewed left, with a crowd skewness of -.23 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by crowd kurtosis of -.96 (SE = .38) and SME kurtosis of -.35 (SE = .38). The nine-raters group's ratings were significantly correlated with ASE ($r = .29$, $p < .01$), as were the SMEs' ($r = .27$, $p < .01$). The difference between the nine-raters group's accuracy and the SMEs' accuracy

was not significant ($p > 0.05$). A paired-samples t-test assessed the mean difference between the nine-raters groups' constructiveness scores (M=3.11, SD = .96) and SME raters groups (M=2.77, SD = .82). There was a significant difference between the two groups; $t(158) = 6.78$, $p < .001$), indicating significantly higher ratings among the crowd raters versus the SMEs.

Six raters. Both tests demonstrated consistency within groups of crowd raters ($\alpha = .83$, ICC(1) = .64) as compared to consistency within groups of SME raters ($\alpha = .92$, ICC(1) = .81). There were no sizable apparent differences between crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, $n_1 = 3$, $n_2 = 6$) = .35, $p > .05$. The six-raters and SME raters groups also exhibited consistency between groups ($r = .76$, $p < .01$); the six-raters group showed good interrater agreement ($r_{wg} = .61$) as compared to SME interrater agreement ($r_{wg} = .79$). Both distributions skewed left, with a crowd skewness of -.23 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by crowd kurtosis of -.95 (SE = .38) and SME kurtosis of -.35 (SE = .38). The six-raters group's ratings were significantly correlated with ASE ($r = .27$, $p < .01$), as were the SME's ($r = .27$, $p < .01$). The difference between the six-raters group's accuracy and the SMEs' accuracy was not significant. A paired-samples t-test assessed the mean difference between the six-raters groups' constructiveness scores (M=3.09, SD = .96) and SME raters groups (M=2.77, SD = .82). There was a significant difference between the two groups; $t(158) = 6.34$, $p < .001$), indicating significantly higher ratings among the crowd raters versus the SMEs.

Three raters. The randomly selected three-raters group did not yield positive reliability statistics ($\alpha = -2.1$; ICC(1) = -4.78). The three-raters and SME raters groups

did, however, demonstrate consistency between groups (r = .76, p < .01). Crowd raters demonstrated moderately similar interrater agreement (r_{wg} = .589) when compared to SME interrater agreement (r_{wg} = .79). Both distributions skewed left, with a crowd skewness of -.34 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by crowd kurtosis of -.72 (SE = .38) and SME kurtosis of -.35 (SE = .38). The three-raters group's ratings were significantly correlated with ASE (r = .28, p < .01), as were the SME's (r = .27, p < .01). The difference between correlations was not significant. A paired-samples t-test assessed the mean difference between the three-raters groups' constructiveness scores (M = 3.16, SD = 1.01) and SME raters groups (M = 2.77, SD = .82). There was a significant difference between the two groups; $t(158) = 6.76$, p < .001).

4.2 Research Question 2

The objective of further analyses conducted for RQ2 was to determine whether filtering raters by certain criteria regarded as common to Subject Matter Experts could improve a group's scores relative to their randomly selected counterparts from RQ1. Given the aim of the present research to support the use of crowd-driven analytical methodologies at an acceptable cost, the results of greatest interest were those of the smallest possible crowd groups to achieve reliability, rather than the results of larger group(s) with the strongest reliability statistics. As the six-raters group was able to demonstrate reliability and accuracy across all indicators as reported for RQ1, and as the RQ1 results for the larger nine- and twelve-raters groups already indicate similar or better performance compared to smaller groups, filtering analyses were not conducted on the larger groups. This approach was additionally supported by the attainment of reliability and accuracy by the twelve-raters, nine-raters, and six-raters groups despite having used

only random selection to determine membership in each group. Furthermore, given the presence of some strengths in the reliability results for the randomly- selected three-raters group, new members were also chosen for these groups based on high levels of certain filtering criteria. Tests were then repeated in order to confirm whether more robust selection methods made results more consistent with larger groups. Descriptive statistics for the Motivation and Engagement scores of these groups are presented in Tables 5 and 6. Results of the selection and elimination filtering analyses are presented in Tables 7a, 7b, and 8, with confidence intervals included for comparison.

Six raters — Motivation selection. After six-raters groups were created using top scores on the Intrinsic Motivation Inventory (IMI; $\alpha = .78$), both reliability tests demonstrated consistency within groups of filtered crowd raters ($\alpha = .86$, ICC(1) = .72) as compared to consistency within groups of SME raters ($\alpha = .92$, ICC(1) = .81). There were no sizable apparent differences between filtered crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, $n_1 = 3$, $n_2 = 6$) = .20, $p > .05$. The filtered six-raters and SME raters groups demonstrated consistency between groups ($r = .72$, $p < .01$) and moderate levels of interrater agreement ($r_{wg} = .59$) when compared to SME interrater agreement ($r_{wg} = .79$). Both distributions skewed left, with a filtered crowd skewness of -.24 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by filtered crowd kurtosis of -.83 (SE = .38) and SME kurtosis of -.35 (SE = .38). Reliability statistics for the filtered six-raters group ($\alpha = .86$, ICC(1) = .72) were both higher than the randomly-selected six-raters group ($\alpha = .83$, ICC(1) = .64). The filtered six-raters' ratings were significantly correlated with ASE ($r = .29$, $p < .01$) as were randomly-selected six-raters' ratings from previous

analyses (r = .27, p < .01) and the SMEs' (r = .27, p < .01). These differences were not significant. A paired-samples t-test assessed the mean difference between the filtered six-raters' constructiveness scores (M=3.14, SD = .98) and SME raters groups (M=2.77, SD = .82). There was a significant difference (.36) between the two groups; $t(158) = 6.81$, $p < .001$).

Six raters — Engagement selection. Both reliability tests demonstrated consistency within groups of filtered crowd raters (α = .86, ICC(1) = .68) as compared to consistency within groups of SME raters (α = .92, ICC(1) = .72). There were no sizable apparent differences between filtered crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, n_1 = 3, n_2 = 6) = .20, $p > .05$. The filtered six-raters and SME raters groups showed similar consistency between groups (r = .75, p < .01) and demonstrated similar interrater agreement (r_{wg} = .67) when compared to SME interrater agreement (r_{wg} = .79). Both distributions skewed left, with a filtered crowd skewness of -.27 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by filtered crowd kurtosis of -.82 (SE = .38) and SME kurtosis of -.346 (SE = .38). As with the motivation-filtered six-raters group, reliability statistics for the engagement-filtered six-raters group were both higher than the randomly-selected six-raters group (α = .83, ICC(1) = .64). The filtered group's ratings were significantly correlated with ASE (r = .27, p < .01) where the randomly-selected group's ratings from previous analyses correlated less strongly (r = .27), as did the SMEs' (r = .27, p < .01). Differences were not significant. A paired-samples t-test assessed the mean difference between the filtered six-raters' constructiveness scores

(M=3.07, SD = .99) and SME raters groups (M=2.77, SD = .82). There was a significant difference (.3) between the two groups; $t(158) = 5.64$, $p < .001$).

Six raters – Experience & level of education selection. Crowd raters' previous experience with the task and level of education were assessed with single-item measures. For education, possible responses were "1 – No high school diploma or degree"; "2 – High school diploma"; "3 – Some college, associate degree"; "4 – College degree, B.A. or B.S."; "5 – Advanced or professional degree." For experience with the task, participants responded from "1 – Lowest" to 5 – Highest" in response to the question, "To what extent do you feel you already have experience with similar tasks [...]?"

When assessed for reliability, neither of these filtering variables returned meaningfully better results as the six-raters random group, especially relative to Motivation and Engagement Selection. While the groups selected using these criteria were similarly distributed in their responses, as with the other groups in the study, the only noteworthy results were between-group consistency for both education ($r = .77$, $p < .01$) and experience ($r = .73$, $p < .01$) and accuracy for both education ($r = .31$, $p < .01$) and experience ($r = .29$, $p < .01$). Given these results, these analyses were not repeated for the three-raters groups.

Three raters – Motivation selection. After three-raters groups were created using high scores on the Intrinsic Motivation Inventory (IMI; $\alpha = .78$), both reliability tests demonstrated consistency within filtered groups of crowd raters ($\alpha = .78$, ICC(1) = .47) as compared to consistency within groups of SME raters ($\alpha = .92$, ICC(1) = .81), an improvement over the results of the randomly-selected three-raters group. There were no

sizable apparent differences between filtered crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups, X^2 (1, n_1 = 3, n_2 = 3) = .48, p > .05. The filtered three-raters and SME raters groups also demonstrated consistency between groups (r = .71, p < .01) and good interrater agreement (r_{wg} = .65) when compared to SME interrater agreement (r_{wg} = .79). Both distributions skewed left, with a filtered crowd skewness of -.21 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by filtered crowd kurtosis of -.99 (SE = .38) and SME kurtosis of -.35 (SE = .38). The filtered three-raters' ratings were significantly correlated with ASE (r = .30, p < .01), as was the randomly-selected three-raters group (r = .30, p < .01) and SMEs' (r = .27, p < .01). Differences were not significant. A paired-samples t-test assessed the mean difference between the filtered three-raters' constructiveness scores (M=3.05, SD = .98) and SME raters groups (M=2.77, SD = .82). There was a significant difference (.28) between the two groups; t(158) = 6.81, p < .001).

Three raters – Engagement selection. After six-raters groups were created using high scores on the task engagement portion of the Short DSSQ, both reliability tests demonstrated high consistency within filtered groups of crowd raters (α = .96, ICC(1) = .89) as compared to consistency within groups of SME raters (α = .92, ICC(1) = .81). There were no sizable apparent differences between filtered crowd and SME raters in terms of within-group consistency, as the alpha statistics did not differ significantly between groups; X^2 (1, n_1 = 3, n_2 = 3) = .23, p > .05. The filtered three-raters and SME raters groups showed similar consistency between groups (r = .67, p < .01). Filtered crowd raters likewise demonstrated moderately similar interrater agreement (r_{wg} = .57)

when compared to SME interrater agreement (r_{wg} = .79). Both distributions skewed left, with a filtered crowd skewness of -.13 (SE = .19) and an SME skewness of -.19 (SE = .19) accompanied by filtered crowd kurtosis of -.97 (SE = .38) and SME kurtosis of -.35 (SE = .38). The filtered three-raters' ratings were significantly correlated with ASE (r = .27, p < .01) where the randomly-selected six-raters' ratings from previous analyses correlated similarly (r = .27), and the SMEs' were identical (r = .27, p < .01). Differences were not significant. A paired-samples t-test assessed the mean difference between the filtered six-raters' constructiveness scores (M=3.14, SD = 1.12) and SME raters groups (M=2.77, SD = .82). There was a significant difference (.37) between the two groups; $t(158) = 5.51$, $p < .001$).

Elimination – Five & four raters. The following analyses employed elimination as previously described as an alternative to top score-based group selection. Participants from the original random group of six were rank-ordered in terms of their scores on each filtering variable. Five-raters groups were then created by eliminating the lowest scoring rater from each of the eight tasks, then four-raters groups were created by eliminating the second lowest. Results are presented by variable below, and presented fully in Tables 9 and 10.

Motivation. Removing one of the least-motivated raters from the six-raters data analyzed to create a five-raters group led to a minimal improvement in results. Removing the two least-motivated raters reduced skewness to -0.16 from -.23, however, making this group one of the least left-skewed in the study and even less left-skewed than the SMEs. Kurtosis saw no similar reduction. r_{wg} slightly increased for the four-raters group (.65 versus .61).

Engagement. Where removing the unmotivated participants marginally improved data quality, albeit insignificantly, removing disengaged participants from the six-raters group decreased interrater agreement with each removal. The decrease in the four-raters group from the randomly selected six-raters group was slight (r_{wg} decreased to .57 from .61), but noteworthy given the high r_{wg} statistic for the SME ratings (.79).

CHAPTER 5: DISCUSSION

Results indicate that, when their work was assessed for reliability and accuracy, randomly selected groups of six raters from the Mechanical Turk user base performed similarly compared to Subject Matter Experts (SMEs). Similar or marginally improved results were reflected in groups of nine raters and groups of twelve raters. Groups using twelve, nine, and six raters were all positively evaluated for both the consistency and agreement dimensions of reliability, for the similarity of their distributions, and for accuracy. Selecting or eliminating specific crowd raters using criteria common to SMEs further improved reliability in groups of six, and subsequently even three, crowd raters relative to randomly selected groups of the same size. The criteria of Engagement and Motivation, particularly Motivation, were shown to have moderate positive effects when used as either selection or elimination criteria for crowdsourced raters, although selection based on top criterion scores impacted more indicators of reliability and similarity to SMEs than did elimination based on lowest criterion scores. One caveat is that the evidence for the statistical significance of these moderate improvements is mixed. At the end of analysis, no crowd group was found to be significantly more or less accurate relative to Subject Matter Experts, regardless of selection method.

A few notable differences were likewise observed, some both significant and consistent, some not significant but nevertheless consistent across all groups analyzed. All crowd groups tested, including those specially selected using either criterion-based selection or elimination, rated passages of constructive self-talk significantly higher than did groups of SMEs, by .28-.37 on a 1-5 point Likert scale. All crowd groups also exhibited similar levels of kurtosis and were consistently lower than SME kurtosis by .5-

.7. When considering some of the common assumptions and properties regarding surveying the general population for data, there are reasons to expect differences like these to emerge. First, as the crowd raters were indeed non-experts, it is possible that normative positive or acquiescent response bias (Cronbach, 1942; Furnham, 1986; Watson, 1992) and tendencies toward extremes (particularly as some participants lost interest moving through the tasks) could partially explain the differences in distribution. Second, the original SME group was expressly encouraged over several weeks to avoid the extreme ends of the scale unless warranted, while the crowd raters were neither encouraged to do the same nor urged not to feel badly for rating self-talk passages poorly. Third and lastly, when introducing the 1-5 constructiveness scale, two broad and oversimplified examples were provided as seen here:

> *Think: How would you rate all of this together?*
>
> *Is this person being highly constructive? That's a 5.*
>
> *Are they not being constructive at all? That's a 1.*

In retrospect, this arguably over-reductive presentation during crowd raters' introduction to key concepts might have unintentionally contributed to the crowd's tendency toward the extremes. Nevertheless, the consistency and relatively similar value of the crowd's kurtosis across all groups and selection procedures bears mentioning.

These results provide a prototypical methodology and instructions that enable groups as small as six crowdsourced raters to provide good quality quantitative ratings of qualitative data for use in research analyses. Findings are consistent with extant research on fuzzy constructs such as political orientation of a text (Hsueh et al., 2009) and natural language tasks (Snow et al., 2008) from disciplines outside of the social sciences. It is

therefore not merely plausible, but demonstrable, for these groups to achieve a level of reliability and accuracy when rating social science constructs that is highly similar, and in many ways statistically indistinguishable, from the ratings of Subject Matter Experts. As few as three crowdsourced raters have the potential to achieve similar levels of quality when rigorous controls on participation are employed. These controls can be implemented on multiple levels, such as by requiring veteran raters filtered by the crowd platform (Snow et al., 2008) and by implementing selection protocols that isolate and prioritize individual differences in motivation and engagement among the raters. This high level of performance was achieved despite that the constructive self-talk construct is a construct that is removed from popular discourse and typically would not be offered to non-researchers or non-practitioners for assessment or ratings tasks.

Concerns regarding data quality from the crowd are not likely to be easily dismissed among some academics, and rightly so. While additional work is needed to fully assuage these concerns, various strategies aimed at the improvement of crowd reliability can have positive effects in crowdsourced content analysis. Both the *selection* of groups based on the *n* most motivated or engaged participants and the *elimination* of a number of participants noticeably improved reliability for six-person groups of crowd raters and helped even three-person groups achieve reliability. Selection generally outperformed elimination, which led to only marginal improvements in data quality compared to the former. Selection based on motivation, in particular, bears future investigation given the consistency of the improvement in reliability across all tests performed. Critically, however, the most broadly replicable tests performed on all groups encompassed testing for similarity of means and distributions, as well as for agreement

and consistency statistics between groups of raters. Inferences regarding the true accuracy of any of the groups were dependent upon the inclusion and, indeed, existence in the archival dataset of a variable predictive of constructiveness in self-talk data.

Even with this caveat, findings are largely consistent with similar discoveries in crowd science (Benoit et al., 2012; Snow et al., 2008; "Zooniverse," 2018) which have suggested the crowd is capable of more robust research-related tasks than previously thought. They are likewise consistent with established wisdom regarding data quality in crowds (Goodman et al., 2013; Hsueh et al., 2009). These authors, as well as the present research, have emphasized the need to ensure (1) that not "just anyone" is permitted into samples intended for crowdsourced content analysis, and (2) that data presented to the crowd for ratings tasks is presented in such a way as to be approachable by members of the crowd. These results have expanded upon this extant literature by demonstrating the crowd's ability to rate less intuitive, more complex, "fuzzier" data in domains of research where such propositions have not been thoroughly examined.

5.1 Implications

The primary theoretical implication of these findings is that the current trend toward deeper investigation into the crowd's true analytical capabilities is fully warranted and could, in fact, be taken farther. To wit, the crowd may be able to provide not just quality data given certain conditions, but provide a robust analytical supplementation, or outright substitution, for an as-yet undetermined proportion of social science research. Furthermore, these findings highlight a growing need for a way to classify types of data in terms of varying levels of accessibility to the crowd, a method of classification which does not currently have any widely-accepted models. The question of what makes

constructive self-talk “fuzzier” than typical information provided or assessed by the crowd, and indeed what makes it similar enough to that same set such that they might be expected to reliably analyze it, remains difficult to answer even in the presence of these results.

The practical significance of these findings is first and foremost in the potential to offload vast quantities of content analysis work, or “coding,” to the general public via the crowd and a small array of carefully arranged quality controls. Content analysis and other similar phases of research tasks related to qualitative data, most of which are traditionally performed by graduate students, ambitious undergraduates, full-time academics, and non-academic researchers, are notoriously time-consuming and could be significantly reduced at a reasonable cost using designs such as those in the present study. Cost per participant is typically considered as the equivalent of local minimum wage for the amount of time expected to complete the task. While such a consideration might not be reasonable for all studies and could indeed be prohibitive for some, those with access to resources adequate for the task could see a marked difference in time to completion. This is particularly true for projects where coding tasks are entirely delegated to the crowd, leaving individuals designing and driving such projects unencumbered by what can often be their most laborious aspect.

Results also show that additional preliminary work, or “pre-screening,” should not, in the majority of cases, be necessary to select crowd participants in advance. Filtering criteria, such as participant motivation, can be quickly and easily assessed in crowdsourced survey respondents as part of the survey, and following their completion of the research task. Moreover, modern crowdsourcing platforms allow data scientists to exclusively

collect data from veteran, or “Master” in the case of M-Turk, crowd raters, as well as also allow crowd raters to select the studies in which they participate based on detailed descriptions provided by the researchers. Both of these components allow for an increased likelihood of quality responses from the outset. After data collection, a crowd participant’s performance on the filtering criteria can be assessed, allowing for streamlined screening processes. This approach does, however, assume that the resources are available to compensate some greater number of raters than the number eventually selected.

5.2 Limitations

As mentioned, the primary limitation at hand was that there existed little consensus from which to work in deeply clarifying the qualifications of constructive self-talk as an appropriately “fuzzy” variable. In other words, there is still much to be determined with regard to what sorts of data, variables, and constructs are accessible to the crowd such that crowdsourced content analyses are useful. The use of constructive self-talk qualifies as a “fuzzier” variable was largely a face-value judgment based on the experience of the researchers involved in this study’s design, and not something that could be thoroughly justified using extant examples from the social sciences literature. This was, perhaps, largely due to the “fuzzy sets” domain of research having originated in the information sciences despite recent steps toward more social and organizational studies. Future studies expressly designed to clarify how degrees of “fuzziness” are differentiated, and to develop a model of classification by which other researchers can easily do so, would help to cement the foundations of this study and future studies in this growing area.

Secondly with regard to limitations, the variables measuring education and level of experience were one-item measures as opposed to the multi-item scales used for motivation and engagement. It is possible that some differences may emerge from more robust analyses of these crowd qualities as potential selection criteria. Thirdly, the presence of an available constructiveness predictor (ASE) as a means to infer accuracy through a strong correlation was a result of ASE's chance inclusion in an archival dataset. Researchers looking to reproduce or expand upon these findings will need to do so using a variable which can arguably support accuracy in a similar way. Variables which correlate more or less strongly might yield particularly useful insights in the replication of these results.

An alternative interpretation of the results also raises question regarding sufficient accuracy among the Subject Matter Expert raters. While the statistical similarity observed between expert and crowd raters of the correlations verifying accuracy can be interpreted encouragingly in tandem with other observed indicators of crowd ratings quality, there is a more sobering read: it can arguably be an indication of poor ratings on the part of the researchers that served as SMEs. Given extant research, it was not a particularly surprising result that crowds, especially at sizes much larger than the archival dataset's SME group, might be comparable or slightly better in terms of their ratings capability. It was, however, somewhat unexpected to see the SME group's ratings' relatively low correlation (r = .27) with Academic Self-Efficacy. The rest of the strong reliability data observed can similarly be interpreted as support for the notion that the SMEs did not submit particularly accurate ratings, internally consistent though they were. If subsequent research were to create more of a consensus around this explanation, the crowd could

have helped us to uncover the extent to which the experts we so often rely upon in research are lacking in actual expertise.

Also regarding the archival dataset, the use of archival SME ratings did not allow for "filtering" analyses used on crowd groups to test for improvements to reliability. While plausible to assert the hypothesis that SMEs submitting such ratings are more highly motivated and engaged as a rule, the absence of such comparisons prevents these results from being able to rule out whether expert ratings improve, and to what degree, relative to the improvements observed in crowd ratings reliability. Having this capability with the archival dataset would have further strengthened the assertion that motivation and engagement were likely responsible for the improvements observed during analyses by showing less improvement in a ratings group already expected to be both motivated and engaged.

While Mechanical Turk does allow for paid control over whether those responding to surveys are "Master" raters, the means of verifying that these raters are actually better quality than average raters on the platform is to trust the platform's server-side mechanisms. While multiple studies have argued for the usage of such qualifications as a way to reduce noise in data, there is presently no way for researchers using such a platform as Mechanical Turk to independently verify that raters are significantly better or more reliable than non-Master raters for a given dataset.

Finally, consistent with past and future studies in this area, the present study was limited by constraints on how many crowd participants the research team could afford to pay. It is certainly the case that select groups even smaller than the overall sample achieved reliability. This being said, the methods used to filter participants based on

ratings quality may have yielded better results given access to a broader sample from which to draw group members.

5.3 Future Research

Aside from the research question-based orientation of this project, and aside from the abundance of caution urged by the aforementioned crowdsourcing researchers with regard to controlling the quality of respondents, there is ample reason to pursue deeper investigation into these findings through broader application and inquiry. Taking into account the preeminent importance of helping researchers better gauge the appropriateness of a dataset for crowdsourced content analysis, the next most important question raised by this research is at what heights in the domain of "fuzziness" the crowd loses its utility. If they can reliably take the place of content raters for certain areas of expertise, and if further refinements continue to cement this capability, it then remains to be demonstrated which qualities actually define Subject Matter Experts in that area of expertise. What are the upper limits of the "citizen scientist" ("Zooniverse," 2018)? And, as evidenced by this projects findings, is an abundance of caution in order before social scientists label a rater an "expert," at least in domains where non-expert crowds have shown competence and reliability?

Studies answering this call would also benefit from broadening their scope beyond the "fuzzy" label itself. True, it is vital for some researchers to understand the extent to which a variable lacks logical boundaries and thus might be too difficult to explain without extensive education in certain disciplines. It is, however, just as important to understand an alternative perspective: that there are variables for which boundaries become logical enough, and at which precision becomes critical enough, that

non-expert raters simply cannot be trusted to reliably (let alone safely) perform the task. The absence of logical boundaries, as just one property of a dataset, might help to qualify some variables which are traditionally, yet incorrectly assumed to be inaccessible to the crowd in disciplines such as psychology and communication studies. Other variables, though fuzzy by definition, might not qualify for the same tests at all if the stakes on rater accuracy are too high. We now know the crowd is capable of research-level tasks in a growing variety of fields, and results such as these show the crowd's seemingly growing capacity to surprise with their accuracy and reliability. Still, a more varied, robust way of determining what level of research with which the crowd can capably assist must go beyond only a handful of measures and predictors.

On the question of more precisely defining expertise in a given field, a wider array of variables common to Subject Matter Experts, beyond motivation and engagement, should be assessed. Level of education and experience can also be tested more precisely than in the present, with the goal of these questions being to determine which predictors outperform the rest in terms of helping crowd raters achieve similar or better performance to SMEs. Indeed, one of the most intriguing aspects of performing these analyses was, and will continue to be, what qualities, controls, filters, and exclusions might help crowd raters actually exceed the performance of traditional research assistants. If this can be definitively addressed in a way that demonstrates accuracy, consistency and broad validity, the extent to which researchers implementing the more time-intensive content analysis approaches could re-allocate their critical faculties for the better could be considerable.

Future research would benefit just as well from more conclusive demonstrations regarding the points at which statistically significant reliability and accuracy differences between crowd and SME ratings emerge. Given the degree and consistency of statistical similarity between expert and non-expert raters groups (in particular, the lack of statistically significant differences in accuracy between crowd and SME raters groups), the data above do not support any conclusions as to which group of raters, SME or crowd, was significantly more or less accurate (or "correct"). This in turn prevented the results from showing which qualities assessed as part of Research Question 2 might have significantly affected accuracy in particular.

Studies attempting a similar design should include language urging participants to understand that low ratings are not "bad ratings," as accurate judgments across the entirety of the relevant ratings scale are essential and there is no need to be wary of "punishing" the authors of self-talk passages with a poor rating. If the findings from this study are replicated with this language added, the likelihood that a positive response bias is responsible for the higher crowd ratings and lower kurtosis would be lessened even if the results were not significantly affected by the change. One caveat is that, by the very same reasoning, there is no way to definitively say that the distribution and mean ratings statistics from the crowd are not more closely aligned with the true scores of the self-talk data than the statistics from the SMEs. Deeper investigation is thus warranted with regard to the consistently higher crowd ratings, especially given the added observation that the crowd's ratings likewise tended to correlate more strongly with ASE than SME ratings across the board (though the differences, again, were ultimately insignificant).

Some aspects of these findings warrant more precise investigation with regard to cost assessments in particular. Simply stated, even for experts, the minimum size of the task is important to achieving measurable reliability. During experimental analyses, it was observed that not even the SME ratings group could achieve reliability if the overall number of ratings assigned for the task was too far below 100. While there was strong consistency between crowd and SME ratings groups with task sizes as small as twenty ratings, acceptable levels of within-group consistency did not emerge until at least sixty ratings had been collected.

This is likely a result of the central limit theorem, which states that data distributions increasingly normalize along with sample size, leading to more reliable outcomes and reducing the likelihood of erratic results emerging from individual samples. Having established that the crowd is capable of coding this fuzzy construct and taking the central limit into account, it follows that an appropriate number of ratings (i.e., adequate size of task) will always be necessary in order for even a group of traditional experts to demonstrate conclusively that their expertise is valuable (where such a demonstration is as important to the study as it was here). So while the number of crowdsourced raters required to provide an approximation of a rating from traditional expertise might be low, the amount of work an individual rater from the crowd needs to perform should be substantial enough that they can provide useful data as part of an overall sample.

5.4 Conclusion

The aim of this project was to demonstrate the potential for modern crowdsourcing techniques as a viable substitute for traditional expertise in the domain

analyzing content for the presence of certain qualitative dimensions. While this manner of crowdsourced expertise is certainly much less plausible in some domains of expertise than in others (rocketry and engineering, to name a few) established social science research strongly suggests the possibility that the crowd is becoming increasingly capable of handling a significant and unprecedented portion of work for the social and organizational sciences. The results of this paper affirm this promising possibility and indicate multiple avenues by which to conduct future research, validate and enrich extant findings around quality controls, and further strengthen the case for widespread, crowdsourced content analysis of complex data.

Table 1

Sample Task Engagement Items From the Short DSSQ

1. "I was determined to succeed on the task."
2. "My attention was directed towards the task."
3. "I felt tired." (reverse-scored)
4. "I felt bored." (reverse-scored)

Table 2

Example of "twelve-raters" crowd ratings merged for analyses (not real data)

Passage	1	2	..	20	21	22	..	40	41	42	..	60	..	160
Rater 1	4	1	..	5	3	2	..	4	1	3	..	3	..	4
Rater 2	3	2	..	4	4	1	..	3	2	4	..	2	..	4
Rater 3	4	1	..	3	3	1	..	3	3	3	..	2	..	3
Rater 4	5	1	..	5	3	2	..	4	2	3	..	3	..	4
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Rater 12	4	2	..	5	4	3	..	4	3	5	..	3	..	4

Table 3

Example of selection vs elimination procedures on a single task group (not real data)

Original Group of 12 *Coder (Criterion)*	Group of 6 Random	Group of 6 Criterion	Group of 5 1 Elim.	Group of 4 2 Elims
1 (2.5)				
2 (4.8)	2 (4.8)	2 (4.8)	2 (4.8)	2 (4.8)
3 (4.2)	3 (4.2)	3 (4.2)	3 (4.2)	3 (4.2)
4 (3.1)				
5 (2.9)	5 (2.9)		5 (2.9)	
6 (4.0)		6 (4.0)		
7 (4.4)		7 (4.4)		
8 (3.6)	8 (3.6)	8 (3.6)	8 (3.6)	8 (3.6)
9 (1.5)				
10 (1.8)	10 (1.8)			
11 (3.5)	11 (3.5)		11 (3.5)	11 (3.5)
12 (4.5)		12 (4.5)		

Table 4

Comparisons of SME ratings groups to randomly selected crowd ratings groups

	SME Raters	Crowd Raters	
	3 Raters	12 Raters	9 Raters
M	2.77 (.82)	3.11 (.98)	3.11 (.96)
Mean Diff. v. SMEs		.34	.34
t (M.Diff. test statistic)		6.67**	6.78**
Reliability			
α (90% CI)	.92 (.76-1)	.8 (.64-.92)	.92 (.84-.97)
ICC(1) (90% CI)	.81 (.41-.99)	.53 (.16-.81)	.82 (.66-.97)
X^2 for α (df, *p*)		.56 (1, .46)	.00 (1, 1.00)
Correlation of SME, Crowd Ratings		.76**	0.76**
Corrected for Unreliability		.89	.83
Mean r_{wg} (90% CI)	.79 (.76-.82)	.65	.63
Accuracy			
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.29** (.17-.41)	.29** (.17-.41)
Corrected for Unreliability	.30	.35	.33
Distribution			
Skewness (90% CI)	-.19 (-.51-.13)	-.27 (-.59-.05)	-.23 (-.55-.08)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.97 (-1.6- -.34)	-.96 (-1.59- -.33)

** p < .01, * p < .05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 4, continued

Comparisons of SME ratings groups to randomly selected crowd ratings groups

	SME Raters	Crowd Raters	
	3 Raters	6 Raters	3 Raters
M	2.77 (.82)	3.09 (.98)	3.16 (1.01)
Mean Diff. v. SMEs		.32	.39
t (M.Diff. test statistic)		6.34**	6.67**
Reliability			
α (90% CI)	.92 (.76-1)	.83 (.62-.96)	—
ICC(1) (90% CI)	.81 (.41-.99)	.64 (.19-.92)	—
X^2 for α (df, *p*)		.35 (1, .55)	—
Correlation of SME, Crowd Ratings		0.76**	0.7**
Corrected for Unreliability		.87	—
Mean r_{wg} (90% CI)	.79 (.76-.82)	.61	.59
Accuracy			
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.27** (.14-.39)	.30** (.17-.41)
Corrected for Unreliability	.30	.31	—
Distribution			
Skewness (90% CI)	-.19 (-.51-.13)	-.23 (-.55- -.09)	-.34 (-.65- -.02)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.95 (-1.58- -.32)	-.72 (-1.35- -.09)

** p < .01, * p < .05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 5

Descriptives for Motivation as a Selection / Elimination Variable, by Group

Group	*n*	*M*	*SD*
6-Raters Random	48	3.77	0.29
6-Raters Selection	48	4.00	0.21
6-Raters Elimination (1)	40	3.85	0.25
6-Raters Elimination (2)	32	3.91	0.23
3-Raters Random	24	3.76	0.30
3-Raters Selection	24	4.15	0.17

Table 6

Descriptives for Engagement as a Selection / Elimination Variable, by Group

Group	*n*	*M*	*SD*
6-Raters Random	48	28.17	3.92
6-Raters Selection	48	29.44	2.53
6-Raters Elimination (1)	40	29.33	2.27
6-Raters Elimination (2)	32	30.13	1.54
3-Raters Random	24	28.17	3.17
3-Raters Selection	24	29.92	2.57

Table 7

SME ratings groups vs 6-person filtered selection crowd ratings groups

	Expert	Crowd…		
	3 Raters	6 Raters (Random)	6 Raters (Motiv.)	6 Raters (Enga.)
M	2.77 (.82)	3.09 (.98)	3.14 (.96)	3.07 (.99)
Mean Diff. v. SMEs		.32	.37	.30
t (M.Diff. test statistic)		6.34**	6.81**	5.64**
Reliability				
α (90% CI)	.92 (.76-1)	.83 (.62-.96)	.86 (.69-.97)	.86 (.69-.97)
ICC(1) (90% CI)	.81 (.41-.99)	.64 (.19-.92)	.72 (.37-.94)	.68 (.28-.93)
X^2 for α (df, *p*)		.35 (1, .55)	.20 (1, .65)	.20 (1, .65)
Correlation of SME, Crowd Ratings		.76**	.72**	.75**
Corrected for Unreliability		.87	.81	.84
r_{wg} (90% CI)	.79 (.76-.82)	.61	.59	.67
Accuracy				
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.27** (.14-.39)	.29** (.17-.41)	.27** (.15-.39)
Corrected for Unreliability	.30	.31	.34	.31
Distribution				
Skewness (90% CI)	-.19 (-.51-.13)	-.23 (-.55- -.09)	-.24 (-.56- -.07)	-.28 (-.59- -.04)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.95 (-1.58- -.32)	-.83 (-1.46- -.2)	-.82 (-1.45- -.19)

** p < .01, * p < 05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 7, continued

SME ratings groups vs 6-person filtered selection crowd ratings groups

	Expert	Crowd…		
	3 Raters	6 Raters (Random)	6 Raters (Edu.)	6 Raters (Exp.)
M	2.77 (.82)	3.09 (.98)	3.08 (1.01)	3.10 (.97)
Mean Diff. v. SMEs		.32	.30	.33
t (M.Diff. test statistic)		6.34**	5.97**	6.07**
Reliability				
α (90% CI)	.92 (.76-1)	.83 (.62-.96)	.43 (-.27-.87)	.69 (.31-.93)
ICC(1) (90% CI)	.81 (.41-.99)	.64 (.19-.92)	—	.38 (-.38-.86)
X^2 for α (df, *p*)		.35 (1, .55)	1.84 (1, .18)	1.00 (1, .32)
Correlation of SME, Crowd Ratings		.76**	.77**	.73**
Corrected for Unreliability		.87	—	.91
r_{wg} (90% CI)	.79 (.76-.82)	.61	.69	.58
Accuracy				
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.27** (.14-.39)	.31** (.18-.42)	.29** (.16-.40)
Corrected for Unreliability	.30	.31	—	.37
Distribution				
Skewness (90% CI)	-.19 (-.51-.13)	-.23 (-.55- -.09)	-.28 (-.6-.04)	-.23 (-.55-.08)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.95 (-1.58- -.32)	-.92 (-1.55- -.29)	-.92 (-1.55- -.29)

** p < .01, * p < 05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 8

SME ratings groups vs 3-person filtered selection crowd ratings groups

	Expert Raters	Crowd Raters		
	3 Raters	3 Raters (Random)	3 Raters (Motiv.)	3 Raters (Enga.)
M	2.77 (.82)	3.16 (1.01)	3.05 (1.29)	3.14 (1.11)
Mean Diff. v. SMEs		.39	.28	.37
t (M.Diff. test statistic)		6.67**	4.58**	5.51**
Reliability				
α (90% CI)	.92 (.76-1)	—	.78 (.33-.99)	.96 (.88-1.0)
ICC(1) (90% CI)	.81 (.41-.99)	—	.47 (-.59-.97)	.89 (.68-1.0)
X^2 for SME vs. crowd α (df, *p*)		—	.20 (1, .65)	.20 (1, .65)
Correlation of SME, Crowd Ratings		.7**	.71**	.67**
Corrected for Unreliability		—	.84	.71
Mean r_{wg} (90% CI)	.791 (.76-.82)	.59	.65	.57
Accuracy				
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.30** (.17-.41)	.30** (.17-.41)	.27** (.14-.38)
Corrected for Unreliability	.27	—	.36	.30
Distribution				
Skewness (90% CI)	-.19 (-.51-.13)	-.337 (.65- -.02)	-.209 (-.52- .11)	-.13 (-.45- .19)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.72 (-1.35- -.09)	-.99 (-1.63- -0.37)	-.97 (-1.6- -.34)

** p < .01, * p < 05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 9

SME ratings groups vs. 6-person elimination-based crowd ratings groups

	Expert Raters	Crowd Raters		
	3 Raters	6 Raters (Random)	5 Raters (6-1 Unmotiv.)	4 Raters (6-2 Unmotiv.)
M	2.77 (.82)	3.09 (.98)	3.12 (1.03)	3.07 (1.06)
Mean Diff. v. SMEs		.32	.34	.30
t (M.Diff. test statistic)		6.34**	6.24**	5.32**
Reliability				
α (90% CI)	.92 (.76-1)	.83 (.62-.96)	.84 (.62-.97)	.84 (.58-.98)
ICC(1) (90% CI)	.81 (.41-.99)	.64 (.19-.92)	.65 (.15-.94)	.61 (-.03-.95)
X^2 for SME vs. crowd α (df, *p*)		.35 (1, .55)	.29 (1, .59)	.27 (1, .61)
Correlation of SME, Crowd Ratings		.76**	.74**	.74**
Corrected for Unreliability		.87	.84	.85
Mean r_{wg} (90% CI)	.79 (.76-.82)	.61 (.58-.64)	.61 (.56-.65)	.65 (.6-.69)
Accuracy				
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.27** (.14-.39)	.27** (.14-.39)	.26** (.13-.39)
Corrected for Unreliability	.30	.31	.31	.30
Distribution				
Skewness (90% CI)	-.19 (-.51-.13)	-.23 (-.55- -.09)	-.24 (-.52- .11)	-.16 (-.47- .16)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.95 (-1.58- -.32)	-.95 (-1.58- -0.32)	-.97 (-1.6- -.34)

** p < .01, * p < 05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

Table 10

SME ratings groups vs. 6-person elimination-based crowd ratings groups

	Expert Raters	Crowd Raters		
	3 Raters	6 Raters (Random)	5 Raters (6-1 Unengag.)	4 Raters (6-2 Unengag.)
M	2.77 (.82)	3.09 (.98)	3.06 (.98)	3.08 (.98)
Mean Diff. v. SMEs		.32	.29	.30
t (M.Diff. test statistic)		6.34**	5.81**	5.83**
Reliability				
α (90% CI)	.92 (.76-1)	.83 (.62-.96)	.79 (.5-.96)	.82 (.53-.98)
ICC(1) (90% CI)	.81 (.41-.99)	.64 (.19-.92)	.57 (-.02-.92)	.61 (.12-.96)
Correlation of SME, Crowd Ratings		.76**	.77**	.75**
Corrected for Unreliability		.87	.91	.86
Mean r_{wg} (90% CI)	.79 (.76-.82)	.61 (.58-.64)	.60 (.56-.64)	.57 (-.04-1.17)
Accuracy				
Correlation of Constructiveness, Self-report Self-Efficacy (90% CI)	.27** (.14-.38)	.27** (.14-.39)	.28** (.15-.39)	.27** (.14-.38)
Corrected for Unreliability	.27	.31	.33	.31
Distribution				
Skewness (90% CI)	-.19 (-.51-.13)	-.23 (-.55- -.09)	-.19 (-.51- .12)	-.22 (-.53- .1)
Kurtosis (90% CI)	-.35 (-.98-.28)	-.95 (-1.58- -.32)	-.97 (-1.6- -0.34)	-.95 (-1.58- -.32)

** p < .01, * p < 05

Note. Motiv. = Motivation; Enga. = Engagement; Edu. = Education; Exp. = Experience

REFERENCES

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Amazon Mechanical Turk. (2018). Retrieved from https://www.mturk.com/worker/help

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1-26.

Benoit, K., Conway, D., Laver, M., & Mikhaylov, S. (2012). Crowd-sourced data coding for the social sciences: Massive non-expert coding of political texts. Presented at the New Directions in Analyzing Text as Data, Harvard University.

Bohannon, J. (2011). Social science for pennies. *Science*, *334*. Retrieved from http://www.uvm.edu/~cdanfort/press/bohannon-science.pdf

Bonabeau, E. (2009). Decisions 2.0: The Power of Collective Intelligence. *MIT Sloan Management Review*, *50*(2), 45–52.

Burlingame, T. (2017, January 12). Pitch Your Best Potato Chip Flavor Idea for Lay's "Do Us A Flavor" Contest. Retrieved from http://www.fritolay.com/blog/blog-post/snack-chat/2017/01/12/pitch-your-best-potato-chip-flavor-idea-for-lay-s-do-us-a-flavor-contest.htm

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2016). Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations. *Journal of Business and Psychology*.

Cho, J. Y., & Lee, E. H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report*, *19*(32), 1.

Creating and Managing Qualifications. (2017). Retrieved from http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/Concepts_QualificationsArticle.html

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33*(6), 401–415.

DeVellis, R. F. (2006). Classical test theory. *Medical care*, *44*(11), S50-S59.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. International Journal of Internet Science, 11, 51–60.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010, June). Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 80-88). Association for Computational Linguistics.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400.

Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC medical research methodology*, *13*(1), 117.

Galloway, A. W. E., Tudor, M. T., & Vander Haegen, W. M. (2006). The Reliability of Citizen Science: A Case Study of Oregon White Oak Stand Surveys. *Wildlife Society Bulletin*, *34*(5), 1425–1429.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224.

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary educational psychology*, *29*(4), 462-482.

Greengard, S. (2011). Following the crowd. *Communications of the ACM*, *54*(2), 20.

Greicius, T. (Ed.). (2018). Multi-planet System Found Through Crowdsourcing. Retrieved from https://www.nasa.gov/feature/jpl/multi-planet-system-found-through-crowdsourcing

Groth, O. (1948). *Die Geschichte der deutschen Zeitungswissenschaft, Probleme und Methoden*. Munich: Konrad Weinmayer.

Gustetic, G., Shanley, L., Benforado, J., & Miller, A. (2014, December 2). Designing a Citizen Science and Crowdsourcing Toolkit for the Federal Government. Retrieved from https://obamawhitehouse.archives.gov/blog/2014/12/02/designing-citizen-science-and-crowdsourcing-toolkit-federal-governmen

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*(3), 399-425.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 27–35.

Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456). ACM.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of consumer research*, *18*(2), 243-250.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology.* London: Sage.

Lippmann, Walter. (1922). *Public opinion.* New York: Macmillan.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley

Matthews, G., Emo, A. K., & Funke, G. J. (2005, July). A short version of the Dundee Stress State Questionnaire. In *Twelfth Meeting of the International Society for the Study of Individual Differences, Adelaide, Australia*.

Martin, H. (1936). Nationalism and children's literature. *Library Quarterly, 6,* 405-418.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, *60*(1), 48-58.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462-472.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Morris, R. (1994). Computerized content-Analysis in management research: A demonstration of advantages and limitations. *Journal of Management*, *20*(4), 903-931.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pope, C., Ziebland, S., & Mays, N. (2000). Analysing qualitative data. *British medical journal*, *320*(7227), 114.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development*, *52*(1), 5-18.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2007). Assessing social presence in asynchronous text-based computer conferencing. *International Journal of E-Learning & Distance Education*, *14*(2), 50-71.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43,* 450-461.

Schank, R. C., & Abelson, R. (1977). *Scripts, goals, plans, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.

Schnurr, P. P., Rosenberg, S. D., & Oxman, T. E. (1992). Comparison of TAT and free speech techniques for eliciting source materials in computerized content analysis. *Journal of Personality Assessment, 58,* 311-325.

Schnurr, P. P., Rosenberg, S. D., & Oxman, T. E. (1993). Issues in the comparison of techniques for eliciting source material in computerized content analysis. *Journal of Personality Assessment, 61,* 337-342.

Simpson, G. E. (1934). *The Negro in the Philadelphia press.* Unpublished doctoral dissertation, University of Pennsylvania.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Computers in Behavioral Science*, (1962), 484–498.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations, 296*.

Waples, D., & Berelson, B. (1941). What the voters were told: An essay in content analysis. *Graduate Library School, University of Chicago*.

Walworth, A. (1938). *School histories at war: A study of the treatment of our wars in the secondary school history books of the United States and in those of its former enemies.* Cambridge, MA: Harvard University Press.

Watson, D. (1992). Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness. *Sociological Methods & Research, 21,* 52–88.

Wilkinson, S., Joffe, H., & Yardley, L. (2004). Qualitative data collection: interviews and focus groups. Sage Publications.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.

Zeldow, P., & McAdams, D. (1993). On the comparison of TAT and free speech techniques in personality assessment. *Journal of Personality Assessment, 60,* 181-185.

Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of medical Internet research, 15*(4), e73.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers, 16*, 417–434.

Zooniverse. (2018). Retrieved from www.zooniverse.org

APPENDIX A: SURVEY LAYOUT

Survey Layout

Page 1

Thank you for taking part in our short survey.
We may not like to admit it, but we all tend to talk to ourselves.
And whether we talk out loud or in our heads, having an "inner monologue" commentating on situations we face in day to day life is very common.

Click "Next >" to continue.

Page 2

Talking to ourselves intentionally and in a helpful, positive way is called "constructive self-talk."

Constructive self-talk is characterized as conveying a rational and nuanced understanding of oneself or a situation, viewing obstacles in the environment as challenges instead of threats, generally including motivational or instructional language, and is usually optimistic (though not naively so).

This is a detailed definition. Please take a moment to consider each of its parts.

In a moment, you will be rating a small group of written self-talk passages. These passages were written by real people, not researchers. Your task is to provide a rating, on a scale of 1 to 5, of the amount of constructiveness you think is present in their self-talk. Therefore, your understanding of the above definition is very important.

We will also need to give you a few practice attempts.

Click "Next >" when you feel ready to begin.

Page 3

In the following practice task, you will be asked to rate three different passages for their constructive self-talk. The responses are real, but you will not know the identity of the respondents.
Ratings will range from 1 to 5 (from "no evidence of constructive self-talk" to "great evidence of constructive self-talk").
If you do not quite understand, think of these ratings as how well the definition of constructive self-talk is being fulfilled by the passage you're rating. Is this person being greatly constructive? That's a 5. Are they not being constructive at all? That's a 1.
Here's the definition again.

Constructive self-talk is characterized as conveying a rational and nuanced understanding of oneself or a situation, viewing obstacles in the environment as challenges instead of threats, generally including motivational or instructional language, and is usually optimistic (though not naively so).

Click "Next >" when you are ready to continue.

Page 4

Remember: Since you'll be using a 5 point scale, a rating of…
- "1" would indicate *no evidence* of constructive self-talk.
- "2" would indicate *little evidence* of constructive self-talk.
- "3" would indicate *some evidence* of constructive self-talk.
- "4" would indicate *good evidence* of constructive self-talk.
- "5" would indicate *great evidence* of constructive self-talk.

Are you ready? The next page will show you the definition one more time. You're welcome to skip past it if you feel confident. Or, you can study it some more.
Then, the following page will begin a brief practice round.

Click "Next >" to continue.

Page 5: W1S2 #177

Please read the following prompt and self-talk response carefully and completely. Remember, your score should reflect the broad / general constructiveness **of the whole response**. Think: How would you rate all of this together? Then, we'll compare your rating to what we thought.

Prompt: Think about an academic challenge that you are currently experiencing (e.g., a difficult class, a hard assignment, etc.). Stop reading and focus on the kinds of thoughts that go through your head when dealing with this challenge for 30 seconds. In a sentence or two, briefly describe the challenge. Next, please write down the unedited dialogue that runs through your mind (i.e., thoughts) when you are thinking about this challenge. Be sure to write in the first person, "I am thinking…" Please write at least a few sentences.

Self-talk response: **I studied a lot for my midterm in one of my classes and still received a poor grade. I started studying a week and half before the test and knew the material well. The test was definitely fair but my teacher graded difficultly. There are only two things graded the rest of the semester so this test will be hard to make up for. It is extremely stressful.**

On a scale of 1 to 5 (no evidence of constructive self-talk to great evidence of constructive self-talk), rate the self-talk response passage.

> ____

Click "Next >" to continue.

Page 6: W1S2 #177 Results

Self-talk response:	**I studied a lot for my midterm in one of my classes and still received a poor grade. I started studying a week and half before the test and knew the material well. The test was definitely fair but my teacher graded difficultly. There are only two things graded the rest of the semester so this test will be hard to make up for. It is extremely stressful.**

Your rating: x
Our rating: y
If the rating is the same or one point off, good job! You can go to the next practice exercise.
If the rating is two points off or more, **please read the prompt again and take a moment to think about why you think your score was off**. Then, you can continue reading.
This passage was low in constructive self-talk because the responder viewed many of their obstacles as threats or as sources of stress. While they were somewhat rational / objective in thinking about their situation, there was no motivational or instructional language. Instead, the responder focused mainly on the problem and lamenting past events they could no longer change.
On the next page, we'll practice again.
Click "Next >" to continue.

Page 7: W1S2 #151

Please read the following prompt and self-talk response carefully and completely. Remember, your score should reflect the broad / general constructiveness **of the whole response**. Think: How would you rate all of this together? Then, we'll compare your rating to what we thought.

Prompt: Think about an academic challenge that you are currently experiencing (e.g., a difficult class, a hard assignment, etc.). Stop reading and focus on the kinds of thoughts that go through your head when dealing with this challenge for 30 seconds. In a sentence or two, briefly describe the challenge. Next, please write down the unedited dialogue that runs through your mind (i.e., thoughts) when you are thinking about this challenge. Be sure to write in the first person, "I am thinking…" Please write at least a few sentences.

Self-talk response: **I have a bibliography for a research paper due next week. I am nervous about it because I know that it will require a lot of work, but I have other things to work on today. As a result I ignoring it to some degree.**

On a scale of 1 to 5 (no evidence of constructive self-talk to great evidence of constructive self-talk), rate the self-talk response passage.

> ____

Click "Next >" to continue.

Page 8: W1S2 #151 Results

Self-talk response: **I have a bibliography for a research paper due next week. I am nervous about it because I know that it will require a lot of work, but I have other things to work on today. As a result I ignoring it to some degree.**

Your rating: x
Our rating: y

If the rating is the same or one point off, good job! You can go to the next practice exercise.
If the rating is two points off or more, **please read the prompt again and take a moment to think about why you think your score was off**. Then, you can continue reading.

This passage was low-to-medium in constructive self-talk because the responder was rationally evaluating their situation and other objectives, and was also straightforward with themselves about their current procrastination. However, there was no instructive or motivational language to help. It is also possible to interpret the responder's plain description of their nervousness and procrastination as more self-defeating than objective and understanding; however, it is okay to interpret it in either way.

On the next page, we'll practice one more time.

Click "Next >" to continue.

Page 9: W1S2 #178

Please read the following prompt and self-talk response carefully and completely. Remember, your score should reflect the broad / general constructiveness **of the whole response**. Think: How would you rate all of this together? Then, we'll compare your rating to what we thought.

Prompt: Think about an academic challenge that you are currently experiencing (e.g., a difficult class, a hard assignment, etc.). Stop reading and focus on the kinds of thoughts that go through your head when dealing with this challenge for 30 seconds. In a sentence or two, briefly describe the challenge. Next, please write down the unedited dialogue that runs through your mind (i.e., thoughts) when you are thinking about this challenge. Be sure to write in the first person, "I am thinking…" Please write at least a few sentences.

Self-talk response: **Within the next 90 days I would like to become a more positive outgoing helpful person that people can look up to. I know it is hard to sometimes put away the drama or troubles I am going through, but I need to remember people are going through their own. Don't think that this is an excuse for you to let others walk all over you though. Stand up for yourself and what you believe in.**

On a scale of 1 to 5 (no evidence of constructive self-talk to great evidence of constructive self-talk), rate the self-talk response passage.
> ____

Click "Next >" to continue.

Page 10: W1S2 #178 Results

Self-talk response:	**Within the next 90 days I would like to become a more positive outgoing helpful person that people can look up to. I know it is hard to sometimes put away the drama or troubles I am going through, but I need to remember people are going through their own. Don't think that this is an excuse for you to let others walk all over you though. Stand up for yourself and what you believe in.**

Your rating: x
Our rating: y

If the rating is the same or one point off, good job! You can go to the next practice exercise.
If the rating is two points off or more, **please read the prompt again and take a moment to think about why you think your score was off**. Then, you can continue reading.

This passage was high in constructive self-talk because the responder has a clear goal in mind and is thinking objectively about the challenges they will face in overcoming it, as well as providing themselves instructive, motivational language that will reinforce their goal-seeking behaviors.

Now that you're done practicing, you are now ready to begin the survey. You will rate 20 passages in total. It is important that you not overthink or second-guess your scores on any responses; please do not rush to judgment, but also, do not linger. Make an honest assessment and go with your first instinct.

Remember, at this point there will no longer be any feedback after each rating.

Click "Next >" to begin the survey.

The following 20 pages will match the training exercises on page 5, 7, and 9 precisely, but with 20 new responses and without feedback pages afterward.

APPENDIX B: ACADEMIC SELF-EFFICACY ITEMS

I am sure about my ability to do my assignments for school.

Compared to others at my school, I think I am good at learning this material.

I am certain I can understand the material presented at my school.

I am sure I can do as well as, or better than, other students at my school on exams.

I am sure I have the ability to understand the ideas and skills taught at my school.

Compared with other students at my school my learning and study skills are strong.

I am certain I can learn the ideas and skills taught at my school.

APPENDIX C: INTRINSIC MOTIVATION INVENTORY (*REVERSE SCORED)

I enjoyed this task very much.

I think I am pretty good at this kind of task.

I put a lot of effort into this task.

It was important to me to do well at this task.

I felt tense while performing at this task.*

I tried very hard while performing at this task.

Performing this task was fun.

I would describe this task as very interesting.

I am satisfied with my performance at this task.

I felt pressured while performing at this task.*

I was anxious while performing at this task.*

I didn’t try very hard at performing this task.*

While performing this task, I was thinking about how much I enjoyed it.

After performing this task for a while, I felt pretty competent.

I was very relaxed while performing this task.

I am pretty skilled at this task.

This task did not hold my attention.*

I couldn’t perform this task very well.*