Spring 2019

The use of crowdsourcing in the development of measurement instruments

Emily Michelle Wetherell
University of Iowa

THE USE OF CROWDSOURCING IN THE DEVELOPMENT OF MEASUREMENT INSTRUMENTS

by

Emily Michelle Wetherell

A thesis submitted in partial fulfillment
of the requirements for the Master of Arts degree in
Psychological and Quantitative Foundations in the
Graduate College of The University of Iowa

May 2019

Supervisor: Professor Stephen Dunbar

ACKNOWLEDGEMENTS

Writing this this thesis has been a lot of fun—It's exciting to execute an idea and share that idea with others. My goal for this thesis was to explore ideas that may help educational researchers save time and money throughout the test development process, and I'm confident that we will have a firm understanding of the usability of crowdsource platform in the years to come. I see this thesis the foundation of a broader research agenda that I can't wait to execute throughout my Ph.D.

This work would not have been possible without the support and guidance of several special people in my life. First, I want to acknowledge my supervisors and colleagues at Iowa Testing Programs. A special thank you is owed to my advisor, Stephen Dunbar, for trusting me to take on such a unique project and supporting my ideas over the past two years. I'm especially grateful for your ability to put things in perspective when I feel like I've bitten off more than I can chew. I also want to recognize my committee members Catherine Welch and Amy Colbert—I greatly appreciate the time you've taken to be a part of this process, and the guidance you have given me throughout my pursuit of this degree. Your flexibility has been more than appreciated.

I truly feel like "thank you," is not enough to acknowledge the support I've received from my family along the way. Mom and Dad, you instilled a love for learning within me at a very young age, and taught me to define success as *my* best performance. You and always reminded me that accomplishing goals takes initiative from myself, not anybody else… Like Dad always says, "The ball is in your court." Beyond your guidance, you've given me three successful siblings as role models. Jill, Aaron, and Caleb, I look to your successes and feel nothing but respect for the work you have all done to get to where you are now. Knowing that you are proud

of me means everything to me. I know it's not a contest (yes it is), but at this time I'd like to officially announce the start of my campaign for the title of 'Most Successful Wetherell Kid." Good luck.

Several others made a key contribution to my making this paper happen. A special mention is owed to Helen Harton, my honors advisor at the University of Northern Iowa. You first introduced me to crowdsourcing as a data collection platform, and it was your advice that led me to a career in Educational Measurement. Prior to working with you, I had no idea I could combing statistics and psychology into a career—So, thank you. Last but not least, the relationships I've built with my fellow graduate students cannot go unmentioned. I especially want to thank my Educational Measurement and Statistics peers Aaron, Catie, Daniela, Juliana, Kayla, and Thapelo for consistently sharing your kindness, knowledge, advice, and friendship with me. Graduate school is challenging, but s*ignificantly* less challenging when you don't have to do it alone (α=.05).

ABSTRACT

Crowdsourcing has gained favor among many social scientists as a method for collecting data because this method is both time- and resource-efficient. The present study uses a within-subject test-retest design to evaluate the psychometric characteristics of crowdsource samples for developing and field testing measurement instruments. As evidenced by similar patterns of psychometric characteristics across time, strong test-retest reliability, and low failure rates of attention check items, the results of this study provide evidence that Amazon Mechanical Turk might represent a fruitful platform for field testing to support the development of a variety of measures. These findings, in turn, have significant implications for resource efficiency in the fields of educational and organizational measurement.

PUBLIC ABSTRACT

Crowdsourcing has gained favor among many social scientists as a method for collecting data because this method is both time- and resource-efficient. The present study evaluates the psychometric characteristics of crowdsource samples for developing and field testing measurement instruments. The results of this study provide evidence that Amazon Mechanical Turk might represent a fruitful platform for field testing to support the development of a variety of measures. These findings, in turn, have significant implications for resource efficiency in the fields of educational and organizational measurement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Crowdsourcing as a method of data collection has gained favor among many social scientists in the past two decades (Litman, Robinson, & Abberbock, 2016; Goodman & Paolacci, 2017; Brabham, 2013). While crowdsourcing as a data collection resource has received much criticism, countless studies have reported its robust utility for acquiring some types of social science data. These studies include significant and optimistic findings about participant attentiveness and feasibility of using crowdsourcing platforms (Goodman & Paolacci, 2017; Hauser & Schwarz, 2016; Goodman, Cryder, & Cheema, 2012).

As crowdsourcing becomes more prevalent in social science research, it becomes equally imperative to ensure that these data collection platforms are not misused. There is evidence that some scientists have begun to use crowdsourcing communities, such as Amazon Mechanical Turk, to collect validity evidence for new measurement instruments (DeSimone, Harms, Vanhove, & Herian, 2015; Rechlin, 2018; Carey, 2016). Although these studies incorporated appropriate research designs, they also serve as evidence for a need to evaluate whether this platform is dependable enough to produce reliable field-test data for purposes of developing new measures. Ultimately, there is little information available that specifically reports the dependability of crowdsource platforms as a tool for researchers seeking to develop and provide evidence for the psychometric characteristics of new research instruments.

Finding resource-efficient methods for developing new measures is imperative as the cost of field testing becomes increasingly prohibitive. Many test development companies, such as Educational Testing Service, have relied upon embedding methods to field test new items (Educational Testing Service, 2019). Other test developers, such as Iowa Testing Programs,

recruit K-12 students to participate in field testing during large-scale testing events. At the same time, it is becoming more common to administer non-cognitive measures such as inventories of personality and motivation, alongside regularly-administered cognitive measures, such as standardized achievement tests (Burnett, Fernandez, Akers, Jacobson, & Smither-Wulsin, 2012). Because of the significant costs of field testing for achievement tests, and the increase in demand for new non-cognitive measures, it will be beneficial to better understand how crowdsource platforms might aid researchers in developing these instruments.

The goal of this thesis is to provide evidence for whether crowdsourced data, also called online panel data, is a viable option for researchers developing new measures (Walter, Seibert, Goering, & O'Boyle, 2018). To accomplish this, an analysis of technical characteristics of items and scales was conducted on a newly developed career interest inventory using a sample collected through Amazon Mechanical Turk.

Statement of Purpose

There is a need to evaluate the extent to which crowdsource platforms supply samples that have sufficient psychometric quality to provide field-test data for new measures. With the current costs of field testing, and ongoing demands for reliable measures, it is imperative to investigate alternative methods for developing these measures. Additionally, the evidence of misuse of crowdsource platforms such as Amazon Mechanical Turk indicate a pressing need for a research study to assess the dependability of those samples.

The purpose of this study is to examine whether crowdsourcing can be an effective tool for researchers developing educational and psychological measures. The primary objective of this study is to investigate the characteristics of item responses collected from workers on Amazon MTurk. A secondary objective of this study is to provide evidence for the internal and

external validity characteristics of the HSEE Interest Inventory, and to compare the psychometric properties of this MTurk sample to a field test sample of 8th grade students. Specifically, the present study examines the following objectives:

1. To evaluate the item response characteristics of Amazon MTurk workers, and determine whether this resource produces a viable sample for field testing and developing a new non-cognitive measure of interests, in the sense of *a* and *b* below.
    a. Determine which respondents are adequately attentive to item instructions
    b. Evaluate the stability or test-retest reliability of a newly developed non-cognitive, career interest inventory
2. To assess the psychometric properties of items and scales of the career interest inventory with respect to reliability and internal/external validity
3. Compare the reliability and validity evidence of a career interest inventory with samples of crowdsource and field test participants
    a. Estimate internal consistency reliability
    b. Compare distributions of item and test responses

Summary

As the use of crowdsourcing for collecting social science data becomes more common, the importance of understanding the proper uses and limitations of these samples cannot be understated. This study evaluates the test-retest reliability of a newly developed non-cognitive interest inventory. Additionally, this study assesses the attentiveness of a single sample of MTurk workers on two data collection occasions. The evidence collected in this thesis is used to assess the dependability of Amazon Mechanical Turk samples for developing new measurement instruments. Chapter II discusses the relevant literature regarding crowdsourcing, Amazon

Mechanical Turk, non-cognitive measures, and career interest inventories. Chapter III describes the methodology used to accomplish the three aforementioned objectives. Chapter IV reports the results of this study, and Chapter V discusses the results and implications of this research. Chapter V also includes suggestions for future research.

CHAPTER II: LITERATURE REVIEW

This chapter describes the relevant literature on developing measurement instruments, crowdsourcing, Amazon Mechanical Turk, non-cognitive measures, and career interest inventories. Given the breadth of each of these topics, only the literature most relevant to the research objectives of the thesis will be discussed. The history of crowdsourcing and the advantages, as well as limitations, of Amazon Mechanical Turk are highlighted. Additionally, the increase in demand for reliable non-cognitive measurement instruments is summarized, and a brief review on career interests and the importance of assessing adolescents' career interests is presented.

Developing Measurement Instruments

The usefulness of any test relies extensively on the validity evidence for that test (Kane, 2013; Messick, 1995). For purposes of educational assessment, psychometricians often develop cognitive achievement tests that may carry high stakes, or significant consequences, associated with a person's performance on that test (Koretz & Hamilton, 2006; Groves, Fowler, Couper, Lepwoski, Singer, & Tourangeau, 2009). However, not all achievement tests have direct consequences associated with a test taker's performance. Beyond the use of achievement tests, educational researchers continuously develop instruments to measure non-cognitive traits. In educational settings, non-cognitive measures assess individuals' general dispositions, such as personality, attitude, or opinion, and typically do not have high stakes associated with scores (Burnett et al., 2012).

The test development process for both cognitive and non-cognitive measures can be time-consuming and resource-intensive. For example, teams of specialists developing a non-cognitive

measure will likely spend a substantial amount of time on the development of test items and the design of field tests for those items. Field testing new items usually requires the recruitment and compensation of students who are willing to spend time outside of school to take tests. As a result, the field test process can be expensive and students willing to participate may not be representative of the total student population in terms of motivation and achievement (Kirkpatrick & Way, 2008). In some instances, the resources spent may not be cost effective in terms of the quality of information acquired during the field test process.

It is imperative that high-stakes tests undergo rigorous and resource-intensive scrutiny throughout the development process. However, educational researchers seeking to explore item statistics in a resource efficient manner may benefit from using an alternative data collection method that is significantly faster and more cost effective than the methods used in traditional field tests (Follmer, Sperling, & Suen, 2017). Additionally, when a non-cognitive measure does not have direct consequences associated with scores, which is often the case, researchers may have more degrees of freedom in determining reasonable ways to evaluate the quality of items and instruments than solely using traditional field test methods.

Crowdsourcing

The term "crowdsourcing" is used in this thesis to describe the approach to data collection that involves leveraging an electronic medium to acquire what has been called online panel data (Walter, et al., 2018). Crowdsourcing is a form of convenience sampling and is not a novel concept for educational and organizational researchers (Farrokhi & Mahmoudi-Hamidabad, 2012). University-based researchers often collect data from undergraduate students at their university, and applied scientists often collect data from paid-for panels or field samples. Around the turn of the twenty-first century, artists and scientists began seeking ideas and

solutions from internet communities, and researchers quickly identified the opportunity these methods held for a new kind of convenience sampling (Brabham, 2013).

According to some sources, the term, "crowdsourcing," was coined in 2006 in a *WIRED* article that explains this novel form of online information collection (Howe, 2006). Because crowdsourcing is an interdisciplinary term, it is often defined and operationalized in differing ways. Brabham (2013) describes several key components that discriminate what *is* and *isn't* crowdsourcing. He provides four essential features: "(1) An organization that has a task it needs performed, (2) a community (crowd) that is willing to perform the task voluntarily, (3) an online environment that allows the work to take place and the community to interact with the organization, and (4) mutual benefit for the organization and the community (Brabham, 2013)."

The umbrella term, "crowdsourcing," can be used to describe a range of data collection entities. These entities are typically online platforms where a "crowd" of people volunteer to perform tasks in exchange for benefits (typically monetary compensation; Brabham, 2013). For example, organizations such as iStockphoto and Getty Images have utilized crowdsourcing methods to monopolize the stock photography market. These platforms serve as communities for stock photographers, and those in need of stock photography, to interact and exchange goods (Howe, 2006).

Another example of a widely-used crowdsourcing platform is InnoCentive (Palfrey, 2011). This platform serves organizations, or "seekers," that have problems which require a specific expertise to solve. Third-party "solvers," can propose solutions to the seekers in exchange for compensation. For example, Colgate-Palmolive was inquiring about a new way to package toothpaste into a tube without emitting fluoride particles into the air, and a physics hobbyist proposed an effective solution that earned him a $25,000 prize (Howe, 2006). Many

large corporations such as Boeing and Procter & Gamble have used InnoCentive to solve skill-specific problems (Brabham, 2013; Palfrey, 2011; Howe, 2006). While the uses of crowdsourcing appear to be unbounded, the present thesis focuses exclusively on research-oriented uses in the context of instrument development.

Crowdsourcing in a Research Context: Amazon Mechanical Turk

Every year over 1,000 published journal articles report using internet-acquired data, and among the most popular of these research-oriented crowdsource platforms is Amazon Mechanical Turk (Landers & Behrend, 2015; Litman, et al., 2016; Goodman & Paolacci, 2017; Brabham, 2013). This platform, nicknamed "MTurk," allows workers to complete requesters' tasks for compensation. Tasks on MTurk are typically those which computers cannot do, but humans can complete with ease: transcribing audio from video files, identifying properties of a photograph, completing surveys, and participating in research studies remotely (Brabham, 2013). Since 2007, over 15,000 published articles have referenced Amazon MTurk (Goodman & Paolacci, 2017).

Amazon MTurk has become the crowdsource platform, or online panel, of choice for researchers for a variety of reasons. First, the flexibility of the platform allows requesters to acquire information at any day and time, and for any price (Goodman & Paolacci, 2017; Buhrmester, Kwang, & Gosling, 2011). This flexibility is beneficial to researchers who are typically limited to collecting data electronically in proctored computer labs or to researchers who have minimal financial resources. Next, Amazon reports that MTurk has over 100,000 active workers (Harms & DeSimone, 2015). This is the largest quantity of available workers of any crowdsource platform. Due to the large sample of workers, requesters have the option to filter only workers who are qualified to complete their tasks (Paolacci, Chandler, & Ipeirotis,

2010). For example, a requester can call for only workers with a college degree, who own a car, and are between ages 25-30. Researchers can also restrict participation to the country in which an Amazon MTurk worker resides. For a comprehensive list of available qualification restrictions, see: https://www.mturk.com/help.

Another useful feature of the MTurk system is that Amazon offers requesters the opportunity to *not* compensate workers who do not complete the tasks with fidelity, or to only seek participation from workers who have a certain success rate (e.g., 95% or higher). Because of these requester options, workers are especially motivated to stay attentive and complete tasks as instructed to maintain their participation qualifications. Finally, MTurk is significantly more cost and time effective than traditional laboratory studies because participants do not need to commute, and researchers do not have to pay for recruitment advertisements (Goodman & Paolacci, 2017). In total, these reasons have motivated thousands of researchers to take advantage of crowdsource data collection (Goodman & Paolacci, 2017).

As with any research method, there are criticisms regarding the quality of data collected through Amazon MTurk. First, MTurk has a very diverse population of workers in terms of education, ethnicity, and native language (Ross, Irani, Silberman, Zaldivar, & Tomlinson 2010). The diversity of the workers is sometimes interpreted by researchers as having increased representativeness, or increased generalizability, however these reputed properties of Amazon MTurk worker population have been questioned (Follmer et al., 2017). Thus, researchers must use caution when making broad interpretations of MTurk data. Next, there is evidence that some Amazon MTurk workers may use MTurk as a principal source of income (Ipeirotis, 2010; Chandler, Mueller, & Paolacci, 2013; Berinsky, Huber, & Lenz, 2012). As a result, these workers may have more exposure to regularly used measures and paradigms than traditional

college student samples, which may desensitize some MTurk participants to instruments frequently used by MTurk researchers (Ipeirotis, 2010).

Lastly, Amazon MTurk participants are often accused of being inattentive, because they are not proctored by a researcher and are motivated principally by money (Chandler et al., 2013). According to Simon (1957) on the theory of cognitive resources, humans tend to use as few cognitive resources as possible when completing a task (Oppenheimer, Meyvis, Davidenko, 2009). Some may assume that participants are not motivated to expend the cognitive resources needed to respond attentively in an un-proctored research environment (Kronsnick, 1991). Although participant attentiveness is a major concern in any research setting, the claim that MTurk workers are inattentive in responding to tasks has been questioned, and evidence to the contrary has been provided by researchers (Hauser & Schwarz, 2016; Goodman, Cryder, & Cheema, 2012). Two methods researchers use to evaluate participant attentiveness are Instructional Manipulation Checks and Attention Checks (Hauser & Schwarz, 2016).

Instructional Manipulation and Attention Checks

Instructional Manipulation Checks (IMC) and Attention Checks (AC) are items that look similar to regular survey items, but whose purpose is to capture an estimate of participants' attentiveness rather than to measure the study variable of interest (Paas, Dolnicar, & Karlsson, 2018; Oppenheimer, et al., 2009). An IMC is used to evaluate whether a participant is attentively following instructions that are part of an experimental manipulation. An AC is used to capture whether participants are paying attention to survey items on a local level.

For example, in a study that manipulates the race of a character in a vignette, an IMC implemented after the vignette may inquire, "What was the race of the woman in the story you just read?" Respondents who fail to identify the race of the character in the vignette can be

identified as inattentive. An example of an AC item might read, “select dislike for this item.” Respondents who do *not* successfully select, “dislike,” can also be identified as inattentive. Inclusion, or exclusion, of data from participants who are inattentive is determined by the researcher. Although some research has addressed issues regarding cleaning data from online sources, at the present time, there is no official standard of practice for inclusion of data from participants who are deemed inattentive (Walter, et al., 2018; DeSimone, Harms, & DeSimone, 2015).

It has been questioned whether Attention Check items can negatively impact the psychometric properties of an inventory. For example, an attention check item can create participants’ awareness that the researcher is evaluating their attentiveness. Some researchers have inquired whether this may lead participants to overthink their responses (Hauser & Schwarz, 2015). Specifically, in instances where the IMC instructs participants to respond counterintuitively, the participant may feel decreased trust for the researcher and modify the response behavior (Hauser & Schwarz, 2015).

It has also been questioned whether embedding an AC or IMC into a questionnaire impacts the psychometric qualities of the questionnaire as a whole. However, Kung, Kwok, & Brown (2018) have contended that embedding an AC or IMC into a questionnaire does not compromise scale validity. Additionally, a meta-analysis by Walter et al. (2018) revealed that using Online Panel Data instead of more traditional data sources does not have a significant effect on the internal consistency of psychology-related measures (see also Porter, Outlaw, Gale, & Cho, 2018).

Amazon Mechanical Turk Demographics

At its inception, Amazon MTurk only allowed persons with US bank accounts to participate. According to Ipeirotis (2010), Mechanical Turk workers at the time were predominantly female (70%), between the ages of 21-35 (54%), and had annual household incomes of less than $60,000 (65%). Ipeirotis (2010) attributed this overrepresentation of female participants in samples of MTurk workers to the samples' abundance of homemakers seeking to supplement their household income.

Since 2010, Amazon has expanded the use of MTurk to more than 60 countries outside of the United States. Worker demographics have shifted slightly as a result of this expansion. One study on the demographics of MTurk workers reports that approximately 47% of workers are from the United States, 34% from India, and 19% from other countries (Ipeirotis, 2010). Workers from India are predominantly male, while workers from the United States are predominantly female. Ross et al., (2010), have found evidence that in the aggregate, approximately 52% of Amazon MTurk workers are female. This indicates that MTurk samples collected more recently may have a gender distribution more similar to that of the general population of the world.

Newer estimates of education and socioeconomic distribution of MTurk workers indicate that the proportion of workers with an advanced degree is increasing, while the average household income of MTurk workers is declining. Though counterintuitive, this shift in demographics can be explained by the substantial number of MTurk workers residing in India who report a household income less than $10,000 (64%; Ross, et al., 2010). Additionally, there is still an overrepresentation of young adult Amazon MTurk workers compared to the population of persons who use the Internet (Ipeirotis, 2010).

As mentioned above, a limitation of Amazon MTurk is that many researchers overestimate the representativeness of MTurk workers to the general population (Follmer, et al., 2017). However, it is possible that for purposes of developing a non-cognitive instrument, the researcher does not necessarily need to know the identity or demographics of participants to assess certain psychometric properties of an item or measurement instrument of interest in the process of field testing.

Developing Measurement Instruments with Amazon MTurk

As previously mentioned, many researchers have used Amazon MTurk to aid in the development of new measures and instruments, but none of these studies cited evidence that this specific crowdsource platform yields data that are dependable enough for such purposes. For example, one study used data collected on Amazon Mechanical Turk to provide reliability and validity evidence for the Five-by-Five Resilience Scale, but the authors cited this platform's effectiveness for surveys as a reason for using it (DeSimone, Harms, Vanhove, & Herian, 2015). In another example, Carey (2018) evaluated the psychometric properties of the Traumatic Grief Scale using Amazon MTurk, again without citing MTurk's dependability for evaluating scale reliability estimates. Finally, in a similar fashion, Rechlin (2016), developed the Mentoring Functions Measure using MTurk without citing evidence for the viable use of this method. While the development of these three scales incorporated thorough and appropriate research designs, these papers highlight the need for the exploration of and justification for using Amazon MTurk for purposes of measurement development.

An emerging issue evident in these recent studies is that Amazon MTurk is known to be useful for purposes of survey administration and marketing research but has yet to be evaluated in terms of usefulness, dependability, and psychometric robustness for developing new measures

that have not already been field tested on traditional samples (Brabham, 2013). Amazon MTurk is potentially a rich resource for data collection in the development of measurement instruments because of the accessibility to a quantity of data at a comparatively low cost to the researcher. However, there is a gap in the present literature regarding whether crowdsourcing is a dependable tool for developing new measurement instruments. This thesis will attempt to shed light on the dependability of the crowdsource platform in question.

Measuring Non-Cognitive Variables

Whereas cognitive measures evaluate skills such as achievement and literacy, non-cognitive measures evaluate intrapersonal dispositions (Burnett et al., 2012). These dispositions include traits such as personality, motivation, general attitudes, self-control, industriousness, perseverance, and interests (Burnett et al., 2012; Rosen, Glennie, Dalton, Lennon, & Bozick, 2010). While exploration of non-cognitive variables in educational setting dates to the 1930's, in recent years there has been an increase in the pairing of non-cognitive measures with cognitive measures due to findings that some non-cognitive variables are positively related to academic outcomes. Understanding students' non-cognitive traits can help educators to create and administer interventions to promote student success (Stankov & Lee, 2014; Burnett et al., 2012).

A key feature of non-cognitive traits is that in some contexts they are assumed to be stable over time (Johnson & Christensen, 2017). Because of the stability of these traits, and their significant predictive ability for education-related outcomes, it may be of interest to measure some non-cognitive variables during adolescence. Specifically, it can be important to evaluate career interests during adolescence because interests are among the most stable personality traits across the lifespan (Xu & Tracey, 2016; Hansen, 2005; Chope, 2011). Additionally, providing insight about the career interests of adolescents provides students with more time to pursue

career-related activities such as internships, job shadow experiences, and relevant coursework (Arrington, 2000). Such information may also be useful in guiding students toward particular types of courses in high school, or in choosing among high schools with particular areas of emphasis. Students who have the opportunity for early career exploration have a matchless advantage for building a résumé or a college application package.

Interest Inventories

An interest inventory is an assessment that evaluates what a person is, and is *not* interested in (Guion, 1965). On a broad level, interest inventory results can provide an individual with an idea of what his or her interests truly are, and can offer suggestions for career-exploration and career-development (Toman & Savickas, 1997; Arrington, 2000; Guion, 1965). Career psychologists define interest as "an attitude toward activities," and there is evidence to support that interests are a key component in motivation to pursue selected activities (Holland, 1997; Guion, 1965).

It must be noted that interest inventories are *not* a measure of skill or abilities, but there is evidence to suggest that skills and self-efficacy play a significant role in the relationship between interest and career choice (Bandura, 2012; Argyropoulou, Sidiropoulou-Dimakakou, & Besevegis, 2007). In other words, an interest inventory may identify what subjects a person may like but cannot identify which skills or activities a given individual will excel at (Guion, 1965; Toman & Savickas, 1997).

Two main purposes for career interest inventories include guidance for career exploration and guidance for personnel selection (Guion, 1965). For example, a student may take an interest inventory to explore which college major s/he may be interested in, whereas an employer may administer an interest inventory to a possible candidate to see if interest or motivation are

sufficient to do the work. This thesis will focus on career interest inventories for the purpose of career exploration.

Interest Evaluation in Adolescence

According to Career Development Theory (Super & Jordan, 1973)., vocational development occurs in a sequence of five stages: Growth (age 1-14), Exploration (age 15-24), Establishment (age 25-44), Maintenance (age 45-64), and Decline (age 65 and on). Students in adolescence find themselves at the end of a growth stage and the beginning of an exploration stage, where they begin to consider their abilities and examine their occupational interests (Super & Jordan, 1973). According to Career Development Theory, adolescence would be an optimal time to begin to measure an individual's career interests.

While there are currently a variety of career interest inventories available for students to take, many of these instruments are costly and time-intensive (Kuder & Findley, 1966; Campbell 1972; American College Testing Program, 2009). Additionally, many of these inventories yield results that provide students with a list of very specific careers that may seem overwhelming for a student who has just begun the career exploration process (Toman & Savickas, 1997). Adolescents with minimal career exploration experience may benefit from a short and reliable inventory that provides an indicator of their level of interest in a given domain of careers.

Summary

This chapter has provided an overview of the research areas relevant to the objectives of this thesis. The suitability of an online data collection environment for field testing and the psychometric characteristics of the resulting data are central to this study. In addition, the extent to which the crowdsourced data are similar to data collected in a more conventional field-testing

environment of a target student population of interest. These topics are discussed in the following chapter.

CHAPTER III: METHODS

This study investigated of the dependability of Amazon MTurk samples for the use of developing new measures, and the resulting evidence is used to describe the psychometric characteristics of the items and scales of an interest inventory. This chapter describes in detail the data collection methods used to examine the following objectives:

1. To evaluate the item response characteristics of Amazon MTurk workers, and determine whether this resource produces a viable sample for field testing and developing a new non-cognitive measure of interests, in the sense of *a* and *b* below.
    a. Determine which respondents are sufficiently attentive to item instructions
    b. Evaluate the stability or test-retest reliability of a newly developed non-cognitive career interest inventory
2. To assess the psychometric properties of items of a career interest inventory with respect to reliability and internal/external validity
3. Compare the reliability and validity evidence of a career interest inventory with samples of crowdsource and field test participants
    a. Estimate internal consistency reliability
    b. Compare distributions of item and test responses

First, the choice to use Amazon Mechanical Turk and the HSEE Career Interest Inventory (Welch & Dunbar, 2017) to accomplish this study's objectives is explained. Next, participants and data collection methods are described in detail. Lastly, the instruments and attention check items used in this study are presented.

Method

This study used a within-subject test-retest design to evaluate the quality of item characteristics from samples collected with Amazon MTurk. This study also provides evidence for the internal and external validity of a career interest inventory. The choice to use this specific interest inventory as the focus of this thesis is a function of both convenience and theoretical foundation. The HSEE Career Interest Inventory was designed by a team that included this researcher and addresses anticipated needs of public and private school jurisdictions to combine interest information with achievement data in matching students to programs of study in high school. Reliability and validity evidence for this measure was collected using a sample of 8th grade students from a large school district in the Midwest. This allows for direct comparability of inventory items between a field test sample and a crowdsource sample. Additionally, the non-sensitive nature of these items make this inventory well suited for field testing via crowdsourcing.

Participants and Cost

Participants for this research study were 500 workers recruited through Amazon MTurk. The Amazon MTurk workers who participated in this study are adults ages 18+ who lived in the United States and actively used their Mechanical Turk worker account at the time of data collection. It must be noted that this method of data collection restricts the participant population to those who know about the platform, are computer-literate, use the internet, and are comfortable pairing a personal bank account with an online account in the Amazon MTurk system.

At Time 1 of this study, the 500 MTurk workers were each paid $.30 for participation. A total of 2 hours and 24 minutes was needed to collect these data, and the total cost was $210

(including MTurk fees). After completing the questionnaire at Time 1, these 500 workers were then qualified to participate at Time 2 but were not required to do so.

The Time 2 administration of the questionnaire occurred over an 8-day timeframe. A total of 327 of the Time 1 participants also participated at Time 2 (65.4%). The cost of the data was $457.80 at Time 2 (including MTurk fees). Although 8 days were needed for the response rate to reach 65%, it is worth noting that the response rate reached 50% within 72 hours. Compensation for Time 2 participation was greater than Time 1 participation because the Time 2 survey was integral to understanding the stability of item characteristics across time.

Procedure

The inventory and other survey questions were administered via a Qualtrics survey limited to the MTurk workplace. To distribute this survey to qualified MTurk workers, a Human Intelligence Task, or HIT in MTurk jargon, was created and made available for workers to select from a list of other available HITs. Participants self-selected to participate in this HIT. According to Vroom's Expectancy Theory, one can assume that participants were more likely to select this HIT among others if they are interested in the subject and believe the amount of compensation is well worth their effort (Vroom, 1964). It is possible that participants with greater interest in the subject were more likely to participate in this HIT. However, the implications that expectancy theory has for subject interest and research participation are similar for both traditional field test and crowdsource data collection, as K-12 field test samples often consist of students who volunteer to participate.

The HIT for the Time 1 survey was titled "Short survey about your interests (<10 minutes)." The HIT for the Time 2 was titled "Short survey about your interests (part two)." Beneath the title of the HIT, workers were presented with a brief description of the study that

states exactly, “Answer brief questions about your interests.” Additionally, the survey was tagged under the *survey*, *short*, *psychology*, *interesting*, *study*, *career*, *interest*, and *research* tags so that qualified workers may access the HIT by searching any of these key words in the HIT search bar. These specific tags were selected based on anecdotal evidence from the primary investigator and another Amazon MTurk worker.

Workers who were interested in participating in this HIT were taken to a page that describes the survey instructions. Figure 1 provides a screen capture of what the HIT looked like to participants. Instructions for this survey read as follows…

> *“We are evaluating the quality of a recently developed career interest inventory at the University of Iowa. We will ask some questions about your interests, education, and career. Use the link below to access the survey. At the end of the survey you will receive a code to paste into the box below to receive credit for taking our survey.* ***Make sure to leave this window open as you complete the survey.*** *When you are finished, you will return to this page to paste the secret code into the box.”*

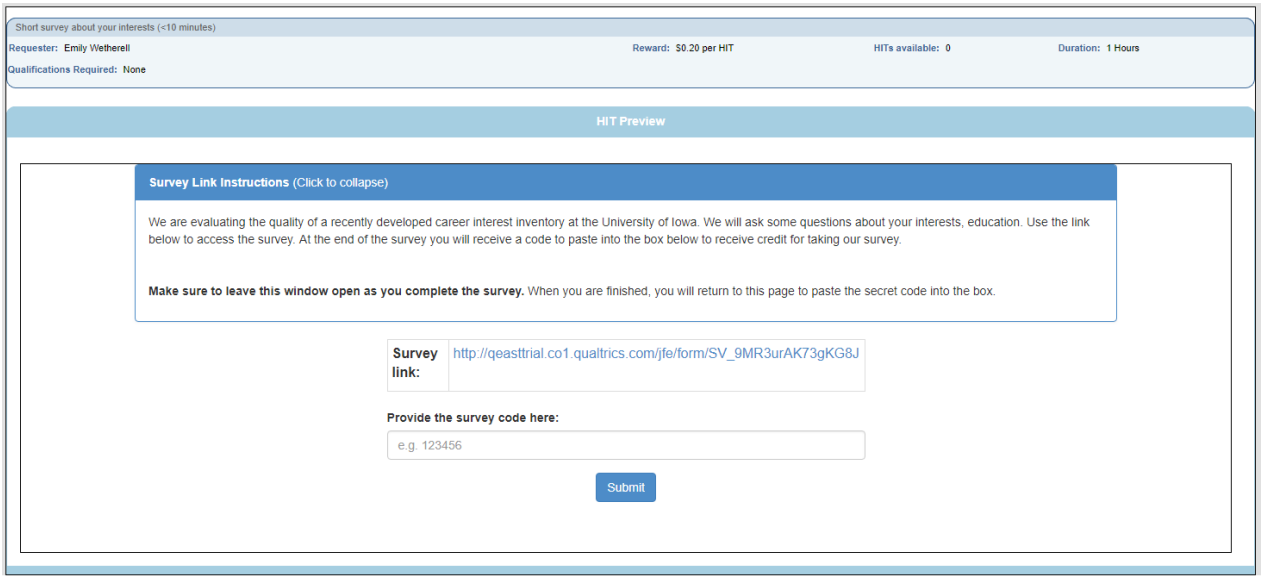


Figure 1. Screen Capture of Amazon MTurk HIT

The Time 1 HIT was released for Amazon MTurk workers to complete on Monday, March 4th, 2019, at 11:00 AM CST. After 500 HITs were completed, the workers were compensated. All 500 workers were compensated regardless of the quality of their responses, because response quality is of great interest for this study. Therefore, responses of low quality are as valuable as responses of high quality in terms of understanding the aforementioned objectives.

On March 11, 2019, at 11:00 AM CST, the Time 2 survey was released to the 500 workers who participated at Time 1. This second survey was released to workers on the same weekday/time as the first portion, under the assumption that most participants who were available to complete the study on Monday at 11:00 AM of a given week might also be available at Monday at 11:00 AM the following week. In order to optimize the participation rate of the second survey, the HIT was made available for participants to complete over the course of an eight-day timeframe. A total of 327 of the 500 potential workers participated at Time 2 (65%). This response rate is consistent with previous findings that follow-up survey response rates with crowdsource samples typically range from 50-60% of the original sample size (Chandler et al., 2013; Daly & Nataraajan, 2015).

Time 1 Design

The survey was designed and administered using the Qualtrics survey distribution system. First, participants were presented with a brief consent-like form (see Appendix A) that describes the purpose and nature of the research study. This form includes information about the study, compensation, participant rights, and potential risks. After reading this information, participants had the option to decide whether they wish to participate by either exiting the screen or clicking an arrow to proceed to the survey items.

Participants were then presented with the HSEE Career Interest Inventory. This inventory has fifty items, although only 49 of these items were administered to the Amazon MTurk sample for reasons described below. The administered inventory is divided into five distinct blocks of items, with four blocks containing 10 inventory items, and one block containing nine inventory items. Each respondent answers one block of the inventory items, then clicks an arrow to proceed to the next page. To prevent artificial inflation of the coefficient alpha reliabilities for the five subscales, items were administered in a spiraled manner such that no two consecutive items correspond to the same career domain (Goodhue & Loiacono, 2002). Additionally, the five item blocks are presented to participants in a random order, as to decrease aggregate primacy and recency response effects (Groves et al., 2009).

After responding to all inventory items, participants had the option to make comments or address any concerns they had regarding the language of the interest inventory. Then, participants self-reported their age by selecting from the following categories: 18-25, 26-35, 36-45, 36-55, 55+. Following this, they were asked questions about their education and current vocation. Figure 2 provides a graphic representation of the questions each respondent was asked after indicating whether they were college.

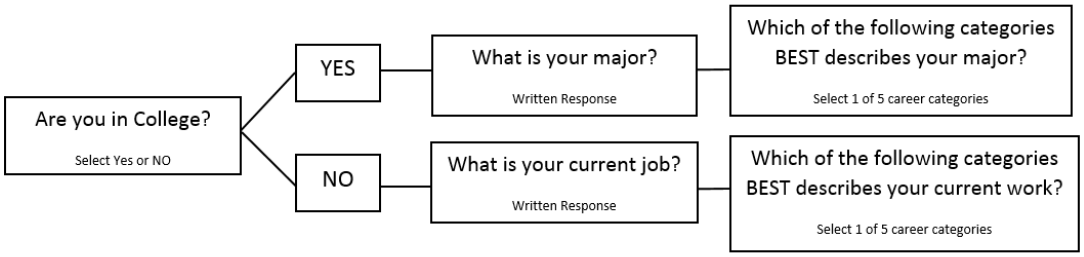


Figure 2. Graphic Representation of Education/Occupation Questions

After completing the education/occupation questions, respondents then self-reported the four-digit year in which they were born. Then, participants were asked to enter their MTurk IDs. This unique ID was necessary to match data between Time 1 and Time 2 of the study. Lastly, respondents were provided with the opportunity to leave any final comments/concerns about the survey. On average, workers took 5 minutes and 13 seconds to complete the Time 1 survey.

Measures

HSEE Career Interest Inventory (Appendix C)

The HSEE Career Interest Inventory was developed because of need for a short, simple, and reliable way to measure 8th grade students' career interests. The results of this inventory are intended to be used with achievement data in matching students to programs of study in high school. The inventory contains 50 items that represent tasks one might do while working in a given career domain. Each item corresponds to one of the following career interest domains: Science/Technology, Fine Arts, Community Services, Business, and Technical/Vocational. These categories were selected based upon review of similar interest inventory sources, and literature concerning the development of vocational interest measures (American College Testing Program, 2009; Holland, 1997). The items were evenly distributed among domains such that each domain had ten corresponding items. Before initial pilot testing, items were evaluated for content fidelity by three test development experts and the primary investigator. A one-hundred percent agreement among all four individuals provided content validity evidence that the items do correspond to the appropriate career domain. Respondents were given the following instructions before responding to the items.

> *"Please indicate whether you would dislike, neither like nor dislike, or like to perform the task. Read each item carefully, and answer honestly. Select responses* ***based on your interests,*** *NOT your skills."*

The items were scored on a three-point scale, including 1=*dislike*, 2=*neither like nor dislike*, and 3=like. The three-point scale was selected because time-efficiency was considered an important factor for the administration of this inventory. Additionally, a primary objective of this research was to obtain stable responses to items, and previous studies indicate that certain psychometric characteristics of a given inventory are independent of the number of Likert-type response options. Thus, a three-point Likert scale was chosen for the purposes of this interest inventory (Jacoby & Matell, 1971). Pilot studies using amazon MTurk yielded internal-consistency reliability estimates (Coefficient alpha) = .76 to .89 (median=.81).

After the first pilot tests were administered, the language of four items was revised to improve clarity, but the meaning of those items remained the same. After revision, the inventory was again pilot tested on a group of 60 workers on Amazon MTurk and yielded somewhat higher internal-consistency reliability estimates (Coefficient alpha) = .80 to .83 (median=.81).

Attention Checks

The first attention check for this survey asked participants to provide a specific response to a question. The item, "Select dislike for this item," is included as the 6th item of one of the five blocks of the interest inventory. Participants fail the attention check if they select *like* or *neither like nor dislike*, or leave the item blank.

Because many researchers who use Amazon MTurk place attention checks in their surveys, it is possible for workers to become familiar with these checks and be exceptionally attentive to ones they recognize (Chandler, et al., 2013). For this reason, a sample that consists of experienced Amazon MTurk workers may yield a pass-rate on attention check items that is artificially inflated. Because of this concern, a unique attention check was included that

compares the consistency of participants' two age-related responses. This attention check is described in greater detail in Chapter IV.

Time 2 Design

As previously mentioned, the Time 2 survey for this study was made available to the 500 Time 1 participants exactly one week after the Time 1 HIT was released. Similar to the Time 1 survey, the Time 2 survey included a consent-like form to provide participants with information about the study, compensation, participant rights, and potential risks (see Appendix B). Then, the participants were presented with the career interest inventory (see Appendix C), and an inquiry of their MTurk ID.

Identical to the Time 1 design, the five blocks of inventory items were administered in a random order to each participant to reduce primacy and recency effects for participant responses. Most participants completed the inventory items in a different order than they did at Time 1. This Time 2 survey did not include the age-related attention check items, but *did* include the "select dislike for this item" attention check. The Time 2 survey also did not inquire about participants' education and occupational background. On average, it took participants took 3 minutes and 59 seconds to complete the Time 2 survey.

For exploratory purposes, as well as to provide additional evidence for internal consistency and participant attentiveness, one item of the interest inventory was administered *twice* in the Time 2 survey. This item ("Protect the public from harm") was embedded in two different item blocks so that participants answered this item on two screens within the same survey. This particular item was inadvertently omitted from Time 1 data collection.

Summary

Although there is some evidence that Amazon MTurk has some utility in social science research with regards to participant attentiveness and resource efficiency, there is little evidence to support the utility of this platform in the development of new measures (Goodman & Paolacci, 2017; Hauser & Schwarz, 2016; DeSimone, Harms, VanHove, & Herian, 2015; Goodman, Cryder, & Cheema, 2012). The within-subject test-retest study presented in this chapter was designed specifically to examine the psychometric characteristics of an MTurk sample for the purposes of evaluating the three aforementioned objectives.

CHAPTER IV: RESULTS

In this chapter, the principal results related to the three research objectives of this thesis are presented. Attention checks, response patterns, and data cleaning are considered first, as they pertain to the response record results included in various data analyses. Results included in this chapter are those deemed critical to specific research objectives, whereas complete summary tables for full documentation of data analyses are included in appendices.

This chapter is organized around item response characteristics (Objective 1), psychometric characteristics of items and scales (Objective 2), and comparisons with traditional field test samples and results (Objective 3). In general, descriptive statistics from various approaches to item and test analyses form the basis for addressing research objectives because the sample sizes obtained provided sufficient stability for purposes of interpretation.

To address the comparison of the Amazon MTurk sample to that of a traditional field test and an existing dataset with responses of a sample of eight grade students (n=133) was used. This sample was administered the interest inventory prior to taking an entrance examination for admission to competitive high schools in a large midwestern city. Analyses of the psychometric properties of the items and scales were replicated on this field test sample to address the comparability of those properties between samples and to address the possibility of using an MTurk sample in lieu of field testing in the same population for which the instrument was intended.

Objective One: Evaluating Item Response Characteristics

Descriptive statistics were calculated to identify how many participants showed evidence of inattentive responses. Response patterns were explored to further ascertain which participants

were responding in a manner that indicated inattentiveness. Additionally, a test-retest analysis examined the relationship between the Time 1 and Time 2 interest inventory responses for this sample of MTurk workers to strengthen the case for data integrity.

Attention Checks

The first objective of this study was to evaluate the item response characteristics of a sample collected from Amazon MTurk. The first steps to accomplish these objectives were to analyze participant responses for attention check items, and to analyze patterns of response that indicate inattentiveness.

The first attention check administered in this study was an item that read "Select Dislike for this item." Participants failed this check if they selected *neither like nor dislike* or *like*, or left the item blank. At the Time 1 administration of the interest inventory, 56/500 (11.20%) of the participants failed this attention check. At the Time 2 administration of the interest inventory, 28/327 (8.56%) of the participants failed this attention check. Of the 28 people who failed this check at Time 2, 11 of them had also failed at Time 1.

The second attention check administered in this study consisted of two age-related questions that were given on different screens of the questionnaire. This check was administered at Time 1 only. The two age-related questions asked respondents to self-report the year they were born, and on another screen to select which of the following categories best described their age: 18-25, 26-35, 36-45, 36-55, 55+. To pass the age-related attention check a respondent must report a birth year that is possible given the response to the age category question. For example, a person who took the inventory on 3/29/2019 and was born in 1993 might be either 25 or 26 years old at the time. Because age 55 was captured twice in these categories, anyone age 55 could select the 36-55 or 55+ category and still pass the attention check.

Descriptive analyses revealed that 16/500 (3.2%) of the Time 1 participants failed the age-related attention check. One participant failed because the characters "Mississi" were given in response to the question about birth year, and another failed because no answer was given to the age category question. The other 14 participants failed because they did not report a birth year that was possible given the response to the age category question. At Time 1, only five participants failed both the *select dislike* and the age-related attention check. This provides an indicator that only a small proportion of MTurk workers were inattentive throughout a single study.

Anomalous Response Patterns

Another method for identifying inattentiveness is observing anomalous patterns of responses to the interest inventory. For example, if a participant selects "dislike" for every inventory item, that participant is exhibiting response behavior that calls into question their attentiveness. In the crowdsource literature, this is often referred to as "invariant responding," or "longstrong responding" (DeSimone, Harms, & DeSimone, 2015). At the Time 1 administration, 8/500 of the participants exhibited this pattern of response. At the Time 2 administration, 5/327 of the participants exhibited this pattern of response. Of the five participants who responded in this manner at Time 2, four of these also responded in this manner at Time 1. It may be possible that participants who exhibit this pattern of response in one survey may do so throughout multiple surveys. Regardless, such response behavior suggests as systematic disregard for the response process.

Data Cleaning

For the purposes of reporting on the integrity of data collected from Amazon MTurk, data from participants who indicated inattentiveness by either failing an attention check, or by

providing a response pattern that indicated inattentiveness, were excluded from further analyses. This yielded a sample of 429 participants at Time 1 and 296 participants at Time 2 for the analyses reported in what follows.

Descriptive Statistics for Interest Inventory Items

Interest inventory items were scored on a scale of 1-3, where 1= *dislike*, 2=*neither like nor dislike*, and 3=*like*. In general, the distribution of item means was similar at Time 1 and Time 2. The means for all interest inventory items at Time 1 ranged from 1.59 (Develop a new smartphone app) to 2.52 (Act in a play or musical). The means for all interest inventory items at Time 2 ranged from 1.67 (Sell real estate to people) to 2.48 (Act in a play or musical/Schedule meetings and appointments). The standard deviations for all items, between both administrations, ranged from 0.73 to 0.91. The means, standard deviations, and item-total correlations (to be described below) for Time 1, Time 2, and the field-test reference sample from the HSEE Technical Manual, are reported in Appendix D.

The items with the highest and lowest means for each domain of the interest inventory are presented in Table 1. It is noteworthy that the items with minimum and maximum means within interest inventory domains were consistent between Time 1 and Time 2 administrations. Because items were administered in 5 randomly ordered blocks, it is unlikely that this pattern is occurred by chance.

Table 1. Min. and Max. Means for Items by Interest Inventory Domain at Time 1 & Time 2

Domain	Item	Time 1 Mean (SD)	Time 2 Mean (SD)
Sci/Tech	Develop a new smartphone app	1.59 (0.81)	1.70(0.86)
	Research diseases to find a cure	2.28 (0.84)	2.31(0.81)
Fine Arts	Sing in a choir or opera	1.82 (0.88)	1.84(0.87)
	Act in a play or musical	2.52 (0.73)	2.48(0.73)
Comm. Services	Talk with people about their problems	1.76 (0.81)	1.83(0.82)
	Serve food to customers at a restaurant	2.29 (0.83)	2.27(0.84)
Business	Sell real estate to people	1.72 (0.83)	1.67(0.81)
	Schedule meetings and appointments	2.46 (0.76)	2.48(0.73)
Tech./Voc.	Fix a broken car engine	1.64 (0.81)	1.75(0.85)
	Install heating & air conditioning equipment	2.33 (0.82)	2.39(0.77)

Similarities among means for interest inventory items between Time 1 and Time 2 were explored using dependent samples t-tests. Means for 43/49 items were not significantly different across time (α=.05). Items whose means differed between Time 1 and Time 2 are reported in Table 2. Although these means were significantly different in a statistical sense, it must be considered that the greatest mean difference ("Teach student in a K-12 classroom") differed by less than two tenths of a standard deviation between Time 1 and Time 2. These statistically significant differences may be a function of large sample sizes rather than effective mean differences. Additionally, means for the item "Protect the public from harm," which was given twice in the Time 2 survey, did not significantly differ from each other (2.17 vs. 2.11). In general, both the results reported in Table 2 and the complete table of means in Appendix D indicate substantial stability in the item means at Time 1 and Time 2 for this sample of MTurk participants.

Table 2. Significant Mean Differences (α=.05) for Items Between Time 1 and Time 2

Item	Mean Difference
Evaluate patient needs and suggest treatments (S2)	-0.098
Write a news story (A1)	-0.094
Teach students in a K-12 classroom (C7)	0.161
Install cables for electrical equipment (V3)	-0.094
Plant, grow, and harvest crops (V7)	-0.124
Fix a broken car engine (V9)	-0.100

To further investigate the stability of responses across time, item-total correlations between item scores and corresponding domain scores were calculated for Time 1 and Time 2. Item-total correlations ranged from .44 to .76 at Time 1, and .45 to .79 at Time 2. In both Time 1 and Time 2 administrations of the interest inventory, each of the 49 items correlated with its corresponding domain at a positive value significantly different from zero (α=.01). In fact, it is possible to show that even the lowest item-total correlation at either time point was significantly greater than .30 using the usual significance tests for correlation coefficients. This provides evidence that the items are hanging together within their respective domains. Items with the highest and lowest item-total correlations for each domain at Time 1 are presented in Table 3.

Table 3. Item-Total Correlations at Time 1

Domain	Item	Correlation
Sci/Tech	Invent a new computer program*	.47
	Conduct an experiment with chemicals	.65
Fine Arts	Act in a play or musical	.49
	Style another person's hair or makeup	.71
Community Services	Talk with people about their problems*	.47
	Plan a special education program for children	.67
Business	Schedule meetings and appointments*	.44
	Manage the activities of other workers	.71
Technical/ Vocational	Develop the landscaping for a home*	.54
	Fix a broken appliance with tools	.76
	Operate heavy machinery such as a forklift or bulldozer	.76

Note: *=same pattern as Time 2

Between the Time 1 and Time 2 administrations, the item with the lowest item-total correlation remained the same for all domains except Fine Arts. For example, the item, "Invent a new computer program" had the lowest item-total correlation in the Science/Technology domain for *both* the Time 1 and Time 2 administrations. Items with the highest and lowest item-total correlations for each domain at Time 2 are presented in Table 4.

Stable results for evidence of item discrimination as illustrated in Table 3 and Table 4 support the coherence of domain scores as well as the utility of the MTurk samples for general test development purposes. Occasion to occasion variability in interest inventory items would beg questions about the stability of self-report measures of career interest. It would also suggest that unexpected aspects of the items chosen to represent a given area of interest cause temporal fluctuations in item response in such as a way that interpretations of scale meaning are suspect.

Table 4. Item-Total Correlations at Time 2

Domain	Item	Correlation
Sci/Tech	Invent a new computer program*	.47
	Design the architecture for a skyscraper	.67
Fine Arts	Write a news story	.49
	Style another person's hair or makeup*	.67
Community Services	Talk with people about their problems*	.48
	Teach students in a K-12 classroom	.63
Business	Schedule meetings and appointments*	.45
	Give financial advice to a client	.68
	Create a budget for an organization	.68
	Interview and hire somebody for a job	.68
Technical/ Vocational	Develop the landscaping for a home*	.50
	Fix a broken car engine	.79

Note: *=same pattern as Time 1

Test Retest Analysis of Domain Scores

A major aspect of the analyses pertaining to the first objective of this thesis is the stability of domain scores from Amazon MTurk over time. A correlation of item scores over time serve as an indicator of the stability of responses provided to items within the questionnaire between occasions. Test-Retest correlations for this study show in Table 5 indicate that this sample produced responses to the interest inventory that were stable between the Time 1 and Time 2 administrations. Moreover, after correcting for measurement error using the coefficient alpha estimates for the five career domains, all test-retest correlations exceed .90. These correlation values, paired with previously mentioned patterns in item means, provide evidence for stability of domain scores between Time 1 and Time 2. Given the spiralized item order and random order of item blocks, it is likely that this result reflects participants' careful responding to inventory items.

Table 5. Interest Inventory Test-Retest Correlations

Domain	Test-retest correlation	Corrected for attenuation
Science/Technology	.74	.95
Fine Arts	.80	.98
Community Services	.75	.98
Business	.79	.97
Technical/Vocational	.80	.91

Note: Coefficient Alpha estimates used to attenuate these correlations are presented in Appendix E

Objective Two: Psychometric characteristics of the HSEE Interest Inventory

The second objective of this research was to assess the internal validity and external validity evidence of a career interest inventory. To accomplish the second objective of this research study, the data from participants at Time 1 who passed the attention checks (n=429) were used. First, descriptive statistics for the five domains of the interest inventory (Science/Technology, Fine Arts, Community Services, Business, and Technical/Vocational) were calculated. Then, inter-domain correlations and coefficient alpha estimates of reliability were computed to evaluate bivariate relationships and the internal consistency of the inventory domain scores. Finally, the relationship between respondents' interest inventory scores and self-reported career or college major was examined.

Descriptive Statistics of Domain Scores

Domain scores for the interest inventory were computed by summing the responses (1=dislike, 2=neither like nor dislike, 3=like) for each domain and dividing by the number of items within the domain. Participants were assigned a score for each of the five domains of the interest inventory: Science/Technology, Fine Arts, Community Services, Business, and

Technical/Vocational. The means and standard deviations of domain scores are reported in Table 6. The mean domain scores ranged from 1.92 to 2.11, and standard deviations ranged from .45 to .59. According to these statistics, it can be inferred that the distribution of responses for interest inventory items were more or less centered, and symmetrical, around a middle value of 2. Appendix F contains graphs of the frequency distributions of scores for the five interest inventory domains.

Table 6. Means and Standard Deviations of Domain Scores at Time 1

Domain Score	Mean	Standard Deviation
Science/Technology	2.03	0.45
Fine Arts	2.11	0.53
Community Services	2.09	0.49
Business	2.09	0.51
Technical/Vocational	1.92	0.59

To further investigate the internal structure of this interest inventory, the intercorrelations among career domains were explored. Pearson correlations for domain scores ranged from .263 (Business & Tech/Vocational) to .655 (Science & Tech/Vocational). After correcting for attenuation, domain correlations ranged from .310 (Business & Technical/Vocational) to .792 (Science & Tech/Vocational). Whereas the observed correlations for domain scores underestimate the true relationships among the domain scores, the disattenuated correlations represent the relationships among the domain scores after accounting for measurement error. The correlations presented in Table 7 suggest that the five domains of the interest inventory are dissimilar enough to represent unique domains of careers yet similar enough to represent subscales of an interest inventory.

Based upon the relationships of similar domains in Holland's theory of vocational choice, this pattern of correlations is not surprising (Holland, Whitney, Cole, & Richards, 1969). Careers such as human resources management (Business), and landscaping (Technical/Vocational) have far less overlap than careers such as architecture (Science/Technology) and landscaping (Technical/Vocational) with respect to interests that may encourage a person to pursue that job.

Table 7. Correlations Among Interest Inventory Domain Scores

	S	**A**	**C**	**B**	**V**
S	-	.52	.61	.44	.78
A	.41	-	.70	.43	.32
C	.46	.55	-	.72	.33
B	.35	.35	.56	-	.28
V	.64	.27	.26	.24	-

Note: All correlations significantly different from zero at α=.01; A=fine arts, C=community services, B=business, V=technical/vocational; below the line are Pearson correlations, above the line are correlations correct for attenuation.

Internal Consistency Reliability

Internal consistency estimates provide a quantitative indicator of how items within a domain of the interest inventory produce responses that are similar to each other. For the Time 1 sample of 429 participants who passed the attention checks, the coefficient alpha values ranged from .76 to .88 (median=.81; See Table 8). Considering the anticipated uses of an interest inventory for high school students, these internal consistency estimates are suitable. It is worth noting that these internal consistency estimates are not exceedingly high, which would serve as an indicator of poor content validity. For example, if the coefficient alpha estimate for Community Services approached 1, the estimate might imply that items within the domain are

not capturing the breadth of possible careers within that domain. Specifically, the Community Services domain encompasses careers that offer services to a community such as teaching, counseling, law enforcement, and working in a restaurant. Different skills and levels of education are required to pursue each of the careers, and a very high coefficient alpha estimate would indicate too narrow a representation of careers within that domain.

Table 8. Internal Consistency (Coefficient Alpha) Estimates for Interest Inventory Domains

Domain	Coefficient Alpha
Science/Technology	.77
Fine Arts	.81
Community Services	.75
Business	.81
Technical/Vocational	.88

Note: includes only participants who were coded as "attentive"

External Validity Evidence

Next, the relationship among interest inventory scores and self-report career or college major was explored. Participants were first assigned a "highest score" based upon their greatest domain score for the interest inventory. For example, if an individual's score on Science/Technology exceeded the scores for the four other domains, that individual was assigned "Science/Technology" as their highest score.

For college students who reported having a major, 22/52 (42.31%) had an exact match between highest interest inventory score and self-reported college major category. For non-college students who reported having a job, 103/311 (33.12%) had an exact match between highest interest inventory score and self-reported career category. A description of the coding scheme used for this analysis is presented in Appendix G.

Objective Three: Comparison of MTurk and Field Test Samples

The third objective of this study was to compare responses to the interest inventory between a crowdsource sample and a traditional field test sample. These comparative analyses used data from the 429 respondents in Time 1 of the MTurk sample that satisfied attention checks, and 133 respondents from a field test sample of 8th grade students. Information on this field test is available in Welch and Dunbar (2018).

Comparison of Descriptive Statistics for Inventory Items and Domains

For the 8th grade sample, item means ranged from 1.44 (Answer telephone calls for a company) to 2.58 (Start your own business), and standard deviations ranged from 0.33 to 0.83. The pattern of highest and lowest average scores for this sample was different than the pattern of highest and lowest average scores for the sample of Time 1 MTurk workers. It is of interest to note that the most disliked item in the Fine Arts domain for both samples was "Sing in a choir or opera." Additionally, "Install heating and air-conditioning equipment" was the most-disliked item of the Technical/Vocational domain for the 8th grade sample, but the most-liked item for the Technical/Vocational domain for the MTurk Time 1 sample (See Table 9).

Table 9. Min. and Max. Means for Items by Interest Inventory Domain (8th grade sample)

Domain	Item	Mean (SD) Time 1
Science/Technology	Perform surgery on a patient	1.85 (0.33)
	Solve advanced math equations	1.85 (0.51)
	Conduct an experiment with chemicals	2.51 (0.51)
Fine Arts	Sing in a choir or opera*	1.55 (0.62)
	Design a logo for a new business	2.42 (0.40)
Community Services	Teach students in a K-12 classroom	1.77 (0.61)
	Respond to people in need during emergencies	2.48 (0.46)
Business	Answer telephone calls for a company	1.44 (0.52)
	Start your own business	2.58 (0.57)
Technical/Vocational	Install heating and air-conditioning equipment	1.48 (0.68)
	Develop the landscaping for a home	2.12 (0.52)

A comparison of mean domain scores between the MTurk Time 1 and Eighth Grade Sample indicates that most of the domain score means are similar across time (See Table 10). Although the difference in Means for the Science/Technology domain differed by less than half of a standard deviation, this difference may be attributable to the eight-grade sample responding differently to specific technology-related items. This assumption is further described in the sections to come. It must also be considered that although the patterns of individual item means vary between samples, the similarities in domain means do not entirely reflect the systematic differences in item means that result as a function of age and career experience. Specifically, there are some items that had high means in one sample, but low means in another (e.g., "Schedule meetings and appointments," "Install heating and air-conditioning equipment," "conduct an experiment with chemicals.")

Table 10. Comparison of Mean Domain Scores Between MTurk and Eighth Grade Samples

Domain	Sample		
	MTurk Time 1 Mean (SD)	Eighth Grade Mean (SD)	Mean Difference
Science/Technology	2.02 (0.49)	2.16 (0.35)	0.144*
Fine Arts	2.10 (0.53)	1.97 (0.49)	-0.135*
Community Services	2.08 (0.48)	2.13 (0.41)	0.048
Business	2.08 (0.51)	2.02 (0.44)	-0.062
Technical/Vocational	1.91 (0.58)	1.81 (0.52)	-0.096

Note: *=significantly different at α=.01

Comparison of Correlations Among Subdomains

For the 8th grade sample, observed correlations among domain scores ranged from -.04 (Fine Arts and Science) to .52 (Community Services and Fine Arts). For the MTurk Sample at Time 1, observed correlations among domain scores ranged from .24 (Business and Technical/Vocational) to .64 (Science/Technology and Technical/Vocational; see Table 12). It is worth noting that in the 8th grade sample, the correlations between the Fine Arts domain and the Technical/Vocational and Science/Technology domain are near zero. This pattern of correlation is dissimilar to the patterns apparent in the MTurk Time 1 sample and may be a function of adolescents' experiences with Fine Arts differing dramatically from adults' experiences with Fine Arts. For example, middle school students are often required to enroll in Fine Arts courses such as band, choir, painting, or ceramics. Their exposure to Fine Arts might occur independent of their interests. In contrasts, adults typically self-select their exposure to the Fine Arts.

Table 11. Correlations Among Domain Scores for MTurk and Eighth Grade Samples

	S	A	C	B	V
S	-	-.04	.22*	.22*	.44*
A	.41*	-	.52*	.22*	-.04
C	.46*	.55*	-	.45*	.16
B	.35*	.35*	.56*	-	.25*
V	.64*	.27*	.26*	.24*	-

Note: *= significantly different from zero at α=.05; A=fine arts, C=community services, B=business, V=technical/vocational; below the line represent correlations for the MTurk Time 1 sample, above the line represent correlations for the 8th grade sample.

Internal Consistency Estimates

To accomplish objective three, it is also of interest to compare the internal consistency reliability estimates for these two samples. As evident in Table 10, the Coefficient alphas for the five domains are similar across all domains of the interest inventory aside from Science/Technology. The younger sample may be responding to the technology-related items in a systematically different way than the Time 1 MTurk sample, because younger generations are assumed have more frequent experiences with technology than older generations. Significance tests for differences of the coefficient alpha statistics in Table 10 (cf. Feldt, Woodruff, & Salih, 1987) may be of interest but are beyond the scope of this presentation of results.

Table 12. Internal Consistency Estimates for Interest Inventory domains for MTurk and Eighth Grade Samples

Domain	Coefficient Alpha	
	MTurk Sample	8th Grade Sample
Science/Technology	.77	.57
Fine Arts	.81	.80
Community Services	.75	.70
Business	.81	.80
Technical/Vocational	.88	.86

Note: this table includes participants who passed all attention checks and did not use a response set; For the Time 1 sample, n=429. For the 8th grade sample, n=133

Summary

This chapter summarized the results for the three objectives of this thesis. Objective 1 explored the psychometric characteristics of and MTurk sample and identified several indicators of quality such as low failure rates for attention check items and stable item means over time. Additionally, strong test-retest correlations support positive findings for this objective. Objective 2 evaluated the psychometric characteristics of an interest inventory and provided quantitative insight into internal and external validity evidence that supports the likely uses of the interest inventory. Finally, Objective 3 compared the psychometric properties of the interest inventory between a sample of MTurk workers and eight grade students. These results indicate that the psychometric properties of the interest inventory differ for these samples in ways that were expected.

CHAPTER V: DISCUSSION

The objectives of this study focused on (1), the item response characteristics of Amazon MTurk responses to questionnaire items, (2) the psychometric properties of items and scales of the interest inventory used in (1), and (3) the comparison of the MTurk sample results with those of a traditional field test sample. This chapter discusses specific findings related to these objectives that lead to interpretations of interest for possible future research.

Objective One: Evaluating item response characteristics

A primary objective of this research was to identify the item response characteristics of a sample of participants from Amazon MTurk. The results of the present study provide several indicators that this sample produced data with psychometric characteristics of sufficient quality for field testing purposes.

First, a significant majority of participants passed both popular and novel attention check items, indicating that *most* participants in this sample were exhibiting effort in responding to the questionnaire items. The proportion of participants who failed attention check items was consistent with findings in previous research (Harms & DeSimone, 2015), which further supports the careful responding of this sample. Beyond that, the differences in item response characteristics were negligible whether inattentive respondents were included or excluded from analyses. Therefore, it may be possible that studies conducted on Amazon MTurk that did *not* include attention checks as a data cleaning method may still be useful with respect to data quality particularly if other characteristics of the data, such as those considered in this thesis, constitute parts of the data quality review.

Several authors have highlighted the issues of data cleaning, especially as it pertains to research via crowdsource platforms. Walter et al. (2018) recommended deleting the first 150 responses from data collected from online panels, and DeSimone, Harms, and DeSimone (2015) recommended cleaning data based upon indicators of psychometric quality and markers of inattentiveness. This avenue of inquiry as it pertains to research on Amazon MTurk and attention check items should be further explored through studies that design specific manipulations to give more complete results on the utility of attention checks. This research might be especially important for evaluating the field-test potential of Amazon MTurk for cognitive measures.

Next, the levels of test-retest reliability for the interest inventory indicate that participants were not responding randomly to the questionnaire and that attentive respondents were the majority on Amazon MTurk. The spiralized item order and randomization of the item blocks in the questionnaire eliminate the possibility that these findings are due to all respondents idiosyncratically responding to the survey on two occasions. This result implies that most participants who were being attentive *within* a single administration of a questionnaire were being attentive *between* different administrations as well.

Third, there was a very low occurrence of invariant or longstring response patterns within these data. For this study, participants were excluded from analyses if they selected the same response for every interest inventory item. Only 8 of 500 participants at Time 1 exhibited this pattern of response. Additionally, at Time 2 only 5 of 327 participants exhibited this pattern of response (Four of these participants had also exhibited this pattern at Time1). These findings provide evidence that an overwhelming majority of participants were *not* using this response pattern and that some workers who *do* use this response pattern in one survey may do so in others. Because the frequency of this pattern occurred at such a low rate, it is difficult to formally

evaluate its impact on the overall quality of samples collected in general on Amazon MTurk. Although a substantial sample size would be necessary to explore this phenomenon, it will be useful for researchers to have a better understanding of invariant response patterns and longstring responses in crowdsource research.

The most substantial indicator of quality of responses from Amazon MTurk was the pattern of item means between the Time 1 and Time 2 administrations of the interest inventory. The items with the minimum and maximum means within each domain remained the same between both administrations. This finding implies that the variation in responses is a result of respondents indicating their actual interests in similar patterns across time. For example, it was not expected that "Sing in a choir or opera" and "Act in a play or musical" would have the lowest and highest means, respectively, within the Fine Arts domain. However, these items maintained the lowest and highest means for that domain in two administrations, which provides some weight to the ranking of these interest items. If respondents were responding carelessly to questionnaires on MTurk, as critics of crowdsourcing may suggest, these patterns of means would not have occurred. Additional methods that take advantage of other summary statistics should be considered in follow-up analyses.

To provide further evidence for the quality of item response characteristics from crowdsource samples, replication of the procedures used in this study with other measures on other crowdsource platforms is warranted. Specifically, the patterns of means for a measure that uses a five or seven-point scale, over time, would be of interest to better understand the psychometric characteristics of crowdsource responses for other survey item types. Such point scales would be expected to display better psychometric qualities with respect to consistency and correlations with external variables. Additionally, a test-retest design of the interest inventory used in this

study with a different online data collection platform would be of interest to further compare the utility of various platforms for purposes of developing new measures.

Whereas the common counterargument against crowdsource data is the inattentiveness and random responding of the workers, these findings provide evidence to the contrary. Taken together, these results highlight the utility that crowdsource platforms may have for educational and organizational researchers developing measurement instruments.

Objective Two: Psychometric characteristics of the HSEE Interest Inventory

The second objective of this study was to provide evidence for the internal and external validity characteristics of the interest inventory. The results from this study indicate that for this sample of MTurk participants, the internal consistency was sufficient for the likely uses of the instrument. Additionally, every interest inventory item correlated strongly with scores for that item's respective domain. These internal consistency indices further support the pattern of correlations between the interest inventory domains, which are consistent with the theoretical patterns outlined by Holland (1997). Together these indicators provide evidence that the interest inventory has coherence within and across the interest domains. This evidence of psychometric quality would not have occurred had the sample responded primarily carelessly to the interest inventory items.

As for external validity evidence, college students demonstrated a higher degree of match between highest inventory score and the category of their self-reported college major than non-college students did. This result may imply that a majority of adults are working in field outside of their college major, and is particularly evidenced by one participant's comment at the end of the survey, "This is interesting. I wish I could do the job based on my interests in real life."

The internal and external validity characteristics of the interest inventory collected from this sample shed some light on the potential uses crowdsourcing may have for the development of measurement instruments in educational and organizational research. Specifically, in the development of educational and organizational measures, MTurk samples may be of utility for exploratory purposes and collecting data for pilot tests. Additionally, crowdsourcing may be a useful tool for researchers developing cognitive measures of achievement. Adult samples from MTurk may be able to identify language issues in items, serve as external consultants for alignment-related studies, and provide insight into rationales for selecting a particular response option. Stable evidence regarding the psychometric properties of the interest inventory used in this study highlight the abundance of opportunities that MTurk offers for test developers to save both time and money.

Objective Three: Comparison of MTurk and Field Test Samples

When comparing the psychometric properties of the MTurk and the 8th grade samples, there is some interest in the extent to which estimates of reliability and validity might be sample-dependent. Additionally, it is important to consider that the career interests of adults and adolescents may differ substantially as a result of lived-experiences and opportunities to explore different careers. A final key consideration when comparing the psychometric properties of the interest inventory between these two samples is the intended purpose of the inventory in each context.

For the field test of 8th grade students, the psychometric properties of the interest inventory were explored for purposes of general guidance and information on possible course selections. For the MTurk study presented in this thesis, the psychometric properties of the interest inventory were explored for purposes of better understanding the integrity of data collected on

Amazon MTurk. In both settings, the psychometric properties of the interest inventory were sufficiently high for the intended purpose. It is expected that the means for items between these samples would differ due to differences in age, lived-experiences, and opportunities to experience different careers. However, it is important to note that although the means for most of the interest inventory items were significantly different between these samples, the means for domain scores were relatively similar. This can be explained by the pattern of mean differences between samples. For the items with statistically different means between samples, approximately half were higher for the eighth-grade sample, and half were higher means for the MTurk Time 1 sample. Additionally, the internal consistency reliability estimates were sufficient for both samples, and item-total correlations within both administrations were greater than .30.

One key finding of interest is that the internal consistency estimates for the domain scores were all similar between these samples, except for the Science and Technology domain. This domain was less internally consistent for the 8th grade sample, which may be attributable to these students coming from the so-called technology generation. Younger people are said to have more exposure to technology, and several items in this domain reference computers, software, and smartphones. Because these participants are assumed to have more exposure to technology, they may view technology that they experience daily in a different way than they view the science they study in school. In contrast, adults may not cognitively separate technology and scientific knowledge as distinctly as adolescents do. This especially evidenced by the significantly higher means for the "Invent a new computer program," "Develop a smartphone app," and, "Conduct an experiment with chemicals," items in the eighth-grade sample.

These differences in the psychometric properties of the items and scales of the interest inventory are sensible upon consideration of the differences in age and career-related experiences

for these samples. However, possibilities for follow-up analyses of the data presented in this thesis include the application of structural equation methods (SEM) to further compare the internal structure of the interest inventory between these samples. For example, the internal structure of the interest inventory could be studied at the item and domain level via confirmatory factor analyses models. Likewise, the Time 1 and Time 2 MTurk, and field test sample, could be considered in an SEM context as multiple-group models.

Conclusion

Of the findings presented in this thesis it important to highlight several key results and implications. First, a substantial majority of the MTurk participants in this study responded to the interest inventory in a way that produced data that represented careful responding both within and between administrations. This is particularly evidenced by the substantial test-retest reliability correlations in this sample, and the relative ranking of item means over time. Second, the internal and external validity evidence for the interest inventory produced by this MTurk sample indicate that these samples respond carefully enough to yield psychometric characteristics that can be supported by theoretical evidence. This opens the door for educational researchers to use MTurk as a source in the exploratory stages in the development of measures. Finally, the MTurk sample produced responses to the interest inventory that were different to the 8^{th} grade in sample in ways that can be theoretically explained by the differences in age and career-related experience between samples. This pattern of differences would not be apparent if the MTurk sample were responding carelessly to items in the interest inventory. Perhaps the most important aspect of findings related to these three objectives is that Amazon MTurk might represent a fruitful platform for field testing to support the development of a variety of educational and organizational psychology measures.

REFERENCES

American College Testing Program, Inc. (2009). *ACT interest inventory technical manual*. Iowa City, IA: Author

Argyropoulou, E. P., Sidiropoulou-Dimakakou, D., & Besevegis, E. G. (2007). Generalized self-efficacy, coping, career indecision, and vocational choices of senior high school students in Greece. *Journal of Career Development, 33*, 316–337. doi:10.1177/0894845307300412

Arrington, K. (2000). Middle grades career planning programs. *Journal of Career Development, 27*, 103–109. doi:10.1177/089484530002700204

Bandura, A. (2012). Social cognitive theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology, Volume 1* (pp. 349-373). London: Sage Publications Ltd.

Berinsky, A. J., Huber, G. A., Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*, 351-368. doi:10.1093/pan/mpr057

Brabham, D. C. (2013). *Crowdsourcing*. MIT Press, MA: Cambridge.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. Retrieved from http://www.jstor.org/stable/41613414

Burnett, S. A., Fernandez, C., Akers, L., Jacobson, J. Smither-Wulsin, C. (2012). Landscape analysis of non-cognitive measures. Retrieved from: https://www.mathematica-mpr.com/download-media?MediaItemId=%7B367D74FE-A269-4164-9BE7-7D5939638B50%7D

Campbell, D. P. (1972). *Handbook for the Strong Vocational Interest Blank*. Stanford, CA: Stanford University Press.

Carey, A. (2018). Death is an amputation: Theoretical development and psychometric validation of a measure of traumatic grief (Master's thesis). Retrieved from https://search-proquest-com.proxy.lib.uiowa.edu/docview/2154855222?pq-origsite=primo

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. doi:10.3758/s13428-013-0365-7

Chope, R. C. (2011). Reconsidering Interests: The next big idea in career counseling theory research and practice. *Journal of Career Assessment, 19*, 343–352. doi:10.1177/1069072710395540

Daly, T. M., & Nataraajan, R. (2015). Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. *Journal of Business Research, 68,* 2603–2609. doi:10.1016/j.jbusres.2015.05.001

DeSimone, J. A., Harms, P. D., Vanhove, A. J., & Herian, M. N. (2015). Development and validation of the Five-by-Five Resilience Scale. *Academy of Management Proceedings, 24*, 778-797 doi:10.5465/ambpp.2015.18417abstract

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36,* 171–181. doi:10.1002/job.1962

Educational Testing Service. (2019). *Computer-delivered GRE® General Test Content and Structure.* Retrieved from https://www.ets.org/gre/revised_general/about/content/computer

Farrokhi, F., & Mahmoudi-Hamidabad, A. (2012). Rethinking convenience sampling: Defining quality criteria. *Theory and Practice in Language Studies, 2,* 784-792. doi:10.4304/tpls.2.4.784-792

Feldt, L. S., Woodruff, D. J., Salih, A. F., 1987. Statistical inference for Coefficient Alpha. *Applied Psychological Measurement, 11*(1), 93-103. doi: 10.1177/014662168701100107

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher*, *46*, 329–334. doi:10.3102/0013189x17725519

Goodhue, D. L., & Loiacono, E. T. (2002). Randomizing survey question order vs. grouping questions by construct: An empirical test of the impact on apparent reliabilities and links to related constructs. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. doi:10.1109/hicss.2002.994385

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224. doi:10.1002/bdm.1753

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210. doi:10.1093/jcr/ucx047

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M, Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Guion, R. M. (1965). *Personnel Testing*. New York, NY: McGraw Hill

Hansen, J.C. (2005). *Assessment of interests*. In S. D. Brown & R. W. Lent (Eds.), Career development and counseling: Putting theory and research to work (pp. 281-304). Hoboken, NJ: John Wiley.

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology, 8,* 183–190. doi:10.1017/iop.2015.23

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. doi:10.3758/s13428-015-0578-z

Howe, J. (2006, June). *The Rise of Crowdsourcing*. Retrieved from: https://www.wired.com/2006/06/crowds/

Holland, J. L. (1997) *Making vocational choices: A theory of vocational personalities and work environments.* Englewood Cliff, NJ: Prentice-Hall.

Holland, J. L., Whitney, D. R., Cole, N.S., & Richards, J. M. Jr. (1969). *An empirical occupation classification derived from a theory of personality and intended for practice and research* (ACT Research Report No. 29). Iowa City, Iowa: ACT, Inc.

Ipeirotis, P. (2010). Demographics of Mechanical Turk. CeDER-10–01 working paper, New York University.

Johnson, B., & Christensen, L. B. (2017). *Educational research: Quantitative, qualitative, and mixed approaches*. Thousand Oaks, Calif: SAGE Publications.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000

Kirkpatrick, R., Way, W. D. (2008). F*ield testing and equating designs for state educational assessments*. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/FieldTestingandEquatingDesignsforStateEducationalAssessments.pdf

Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education/Praeger.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive* Psychology, 5, 213–236. doi:10.1002/acp.2350050305

Kuder, G. F., & Findley, W. G. (1966). The Occupational Interest Survey. *The Personnel and Guidance Journal, 45,* 72–77. doi:10.1002/j.2164-4918.1966.tb03070.x

Kung, F. Y. H., Kwok, N., & Brown, D. (2017). are attention check questions a threat to scale validity? *Applied Psychology: An International Review, 67,* 264-283. doi:10.31234/osf.io/9vrdn

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology, 8,* 142–164. doi:10.1017/iop.2015.13

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49,* 433–442. doi:10.3758/s13428-016-0727-z

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology, 56,* 506–509. doi:10.1037/h0033601

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry to score meaning. *American Psychologist, 50,* 741-749. doi:10.1037//0003-066x.50.9.741

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. doi:10.1016/j.jesp.2009.03.009

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional Manipulation Checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing, 35,* 258–269. doi:10.1016/j.ijresmar.2018.01.003

Palfrey, J. (2011). *Intellectual Property Strategy*. MIT Press, MA: Cambridge.

Paolacci, G., Chandler, J., Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411-419. Retrieved from: http://sjdm.org/~baron/journal/10/10630a/jdm10630a.pdf

Porter, O. L. H., Outlaw, R., Gale, J. P., & Cho, T. S. (2018). The Use of Online Panel Data in Management Research: A Review and Recommendations. *Journal of Management, 45*(1), 319–344. doi:10.1177/0149206318811569

Rechlin, A. M., (2016). The development and validation of the mentoring functions measure (Master's thesis). Retrieved from https://search.proquest.com/docview/1832310765?pq-origsite=primo

Rosen, J., Glennie, E., Dalton, B., Lennon, J., & Bozick, R. (2010). Noncognitive Skills in the Classroom: New Perspectives on Educational Research. doi:10.3768/rtipress.2010.bk.0000.1009.4

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In CHI EA '10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, pp. 2863–2872, New York, NY, USA. ACM

Simon, H. A. (1957). *Models of man: Social and rational.* New York: John Wiley and Sons Inc..

Stankov, L., & Lee, J. (2014). Quest for the best non-cognitive predictor of academic achievement. *Educational Psychology, 34*(1), 1–8. doi:10.1080/01443410.2013.858908

Super, D. & Pierre, J. J. (1973). Career development theory. *British Journal of Guidance and Counselling. 1*(1). 3-16. doi:10.1080/03069887300760021.

Toman, S. M., & Savickas, M. L. (1997). Career choice readiness moderates the effects of interest inventory interpretation. *Journal of Career Assessment, 5*, 275–291. doi:10.1177/106907279700500302

Vroom, V. H., (1964). *Work and Motivation*. New York: Wiley, 1964

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2018). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology*. Advance online publication. doi:10.1007/s10869-018-9552-y

Welch, C., & Dunbar, S. B., (2017). *High School Entrance Examination technical manual*. Iowa City, Iowa: Iowa Testing Programs.

Xu, H., & Tracey, T. J. G. (2016). Stability and change in interests: A longitudinal examination of grades 7 through college. *Journal of Vocational Behavior, 93*, 129–138. doi:10.1016/j.jvb.2016.02.002

APPENDIX A
Study Description for Time 1

We invite you to participate in part one of a two-part research study being conducted by investigators from The University of Iowa. The purpose of the study is to evaluate a newly developed career interest inventory.

If you agree to participate, we would like you to complete a survey about your interests. After this, you will be asked to provide demographic information including age, occupation, or college major: You are free to skip any questions that you prefer not to answer. It will take approximately 10 minutes.

As part of your participation you will be compensated $0.30. You will also become qualified to complete a future part-two of this research study.

We will not collect your name or any identifying information about you. However, at the end of the survey you will be asked to provide your Amazon MTurk ID to be qualified for participation in part-two of this study. Choosing to not provide your MTurk ID will not affect your compensation. IDs will be deleted from record immediately after data analysis. Summarized data may be published in a scholarly or academic setting.

Taking part in this research study is completely voluntary. If you do not wish to participate in this study, please return the HIT before answering any survey questions.

If you have questions about the rights of research subjects, please contact the Human Subjects Office, 105 Hardin Library for the Health Sciences, 600 Newton Rd, The University of Iowa, Iowa City, IA 52242-1098, (319) 335-6564, or e-mail irb@uiowa.edu. You may also contact the primary investigator of this study, Emily Wetherell, at Emily-Wetherell@uiowa.edu

Thank you very much for your consideration of this research study.

If you have read the information above and wish to participate in this survey, please click the arrow to continue.

APPENDIX B
Study Description for Time 2

We invite you to participate in part two of a two-part research study being conducted by investigators from The University of Iowa. The purpose of the study is to evaluate a newly developed career interest inventory.

If you agree to participate, we would like you to complete a survey about your interests: You are free to skip any questions that you prefer not to answer. It will take approximately 10 minutes.

As part of your participation you will be compensated $1.00.

We will not collect your name or any identifying information about you. However, at the end of the survey you will be asked to provide your Amazon MTurk ID. Choosing to not provide your MTurk ID will not affect your compensation. IDs will be deleted from record immediately after data analysis. Summarized data may be published in a scholarly or academic setting.

Taking part in this research study is completely voluntary. If you do not wish to participate in this study, please return the HIT before answering any survey questions.

If you have questions about the rights of research subjects, please contact the Human Subjects Office, 105 Hardin Library for the Health Sciences, 600 Newton Rd, The University of Iowa, Iowa City, IA 52242-1098, (319) 335-6564, or e-mail irb@uiowa.edu. You may also contact the primary investigator of this study, Emily Wetherell, at Emily-Wetherell@uiowa.edu

Thank you very much for your consideration of this research study.

If you have read the information above and wish to participate in this survey, please click the arrow to continue.

APPENDIX C

HSEE Career Interest Inventory Items and Operational Definitions of Career Domains

Science & Technology- This domain of occupation describes careers that involve the extensive use of skills related to Science, Technology, Engineering, and Mathematics. Additionally, this domain includes medical professions and other occupations that are driven by research, evidence, and exploration. Examples include: Doctor, engineer, nurse, architect, physical therapist, research scientist, IT personnel, software engineer, &occupational therapist.

1. Research diseases to find a cure
2. Evaluate patient needs and suggest treatments
3. Solve advanced math equations
4. Conduct an experiment with chemicals
5. Perform surgery on a patient
6. Invent a new computer program
7. Fix a broken electronic device
8. Provide care to patients staying in a hospital
9. Develop a new smartphone app
10. Design the architecture for a skyscraper

Fine Arts- This domain of occupation describes careers that involve the extensive use of creative skills including performing, creating, designing, and describing. Example careers include: Artist, actor/actress, author, web designer, photographer, musician, model, & interior designer.

1. Write a news story
2. Perform music for a live audience
3. Sing in a choir or opera
4. Design clothes for models to wear on a runway
5. Style another person's hair or makeup
6. Write a book or a play
7. Act in a play or a musical
8. Design a logo for a new business
9. Take professional photographs
10. Decorate a client's home

Community Services- This domain of occupation describes careers that involve the extensive use of communication and helping skills. Additionally, this domain includes teaching and non-profit professions. Example include: K-12 teacher, police officer, librarian, therapist, & detective.

1. Protect the public from harm
2. Respond to people in need during emergencies
3. Provide evidence in a courtroom
4. Plan a special education program for children
5. Organize a fundraiser
6. Talk with people about their problems
7. Teach students in a K-12 classroom

8. Serve food to customers at a restaurant
9. Help people who have disabilities with every-day tasks
10. Collect evidence to show how a crime was committed

Business- This domain of occupation describes careers that involve the extensive use of business skills such as management, leadership, administration, personnel selection, & sales. Examples include: Marketing specialist, administrative assistant, product manager, accountant, financial adviser, realtor, etc.

1. Interview and hire somebody for a job
2. Give financial advice to a client
3. Manage the activities of other workers
4. Create a budget for an organization
5. Start your own business
6. Buy, sell, and trade stocks
7. Decide how a company should advertise its products
8. Sell real estate to people
9. Schedule meetings and appointments
10. Answer phone calls for a company

Technical & Vocational- This domain of occupation describes careers that involve the extensive use of technical and hands-on skills such as constructing, repairing, and building. Examples include: electrician, mechanic, plumber, carpenter, farmer, & welder.

1. Develop the landscaping for a home
2. Operate heavy machinery such as a forklift or a bulldozer
3. Install cables for electrical equipment
4. Fix a broken appliance with tools
5. Install heating and air-conditioning equipment
6. Build furniture, cabinets, or other structures out of wood
7. Plant, grow, and harvest crops
8. Build and repair powerlines or wind turbines
9. Fix a broken car engine
10. Use a torch to mold structures out of metal

APPENDIX D

Descriptive statistics for Interest Inventory Items across MTurk samples and 8th grade student sample

Table A.1 Means, Standard Deviations, and Item-Total Correlations for Interest Inventory Items

	Item Task	MTurk Sample Time 1			MTurk Sample Time 2			8th grade sample		
		Mean	SD	Item-Total Correlation	Mean	SD	Item-Total Correlation	Mean	SD	Item-Total Correlation
S1	Research diseases to find a cure	2.28	0.84	0.61	2.32	0.81	0.62	2.48*	0.33	0.33
S2	Evaluate patient needs and suggest treatments	2.04	0.87	0.57	2.09*	0.85	0.55	2.20	0.50	0.50
S3	Solve advanced math equations	1.85	0.88	0.48	1.94	0.87	0.48	1.85	0.33	0.33
S4	Conduct an experiment with chemicals	2.03	0.88	0.65	2.08	0.86	0.66	2.51*	0.39	0.39
S5	Perform surgery on a patient	2.12	0.88	0.57	2.12	0.84	0.57	1.85*	0.51	0.51
S6	Invent a new computer program	1.94	0.86	0.47	1.95	0.85	0.47	2.11*	0.61	0.61
S7	Fix a broken electronic device	2.25	0.84	0.61	2.24	0.86	0.64	2.02*	0.47	0.47
S8	Provide care to patients staying in a hospital	2.00	0.88	0.60	2.03	0.85	0.64	2.29*	0.43	0.43
S9	Develop a new smartphone app	1.59	0.81	0.53	1.70	0.86	0.60	2.20*	0.55	0.55
S10	Design the architecture for a skyscraper	2.17	0.88	0.59	2.19	0.89	0.67	2.14	0.40	0.40
A1	Write a news story	2.14	0.86	0.54	2.22*	0.85	0.49	1.77*	0.43	0.43
A2	Perform music for a live audience	1.95	0.91	0.66	2.01	0.89	0.62	1.96	0.52	0.52

Table A.1—Continued

A3	Sing in a choir or opera	1.82	0.88	0.70	1.84	0.87	0.66	1.55*	0.62	0.62
A4	Design clothes for models to wear on a runway	1.98	0.88	0.65	1.95	0.84	0.62	1.87	0.68	0.68
A5	Style another person's hair or makeup	2.07	0.91	0.71	2.06	0.89	0.67	1.90	0.72	0.72
A6	Write a book or a play	2.24	0.83	0.57	2.27	0.81	0.57	1.87*	0.59	0.59
A7	Act in a play or a musical	2.52	0.73	0.49	2.48	0.73	0.55	1.78*	0.65	0.65
A8	Design a logo for a new business	2.19	0.87	0.56	2.21	0.84	0.65	2.42*	0.40	0.40
A9	Take professional photographs	1.9	0.89	0.61	1.91	0.87	0.58	2.29*	0.76	0.76
A10	Decorate a client's home	2.27	0.85	0.56	2.33	0.81	0.58	2.22	0.61	0.61
C2	Respond to people in need during emergencies	2.11	0.84	0.59	2.08	0.86	0.61	2.48*	0.46	0.46
C3	Provide evidence in a courtroom	2.02	0.86	0.57	2.13	0.84	0.59	2.15	0.40	0.40
C4	Plan a special education program for children	2.12	0.85	0.67	2.09	0.82	0.61	2.02	0.69	0.69
C5	Organize a fundraiser	2.03	0.87	0.61	1.96	0.86	0.60	2.14	0.71	0.71
C6	Talk with people about their problems	1.76	0.81	0.47	1.83	0.82	0.48	2.19*	0.57	0.57
C7	Teach students in a K-12 classroom	2.18	0.84	0.64	2.06*	0.84	0.63	1.77*	0.61	0.61
C8	Serve food to customers at a restaurant	2.29	0.83	0.45	2.27	0.84	0.50	1.95*	0.51	0.51
C9	Help people who have disabilities with every-day tasks	2.04	0.87	0.55	2.05	0.83	0.58	2.02	0.61	0.61

Table A.1—Continued

C10	Collect evidence to show how a crime was committed	2.28	0.82	0.62	2.22	0.82	0.64	2.47*	0.33	0.33
B1	Interview and hire somebody for a job	2.16	0.84	0.65	2.10	0.83	0.68	2.16	0.64	0.64
B2	Give financial advice to a client	1.99	0.87	0.68	2.04	0.85	0.68	1.81*	0.57	0.57
B3	Manage the activities of other workers	2.1	0.84	0.71	2.07	0.82	0.67	2.21	0.61	0.61
B4	Create a budget for an organization	2.11	0.83	0.63	2.10	0.83	0.68	1.86*	0.72	0.72
B5	Start your own business	2.15	0.85	0.61	2.20	0.81	0.56	2.58*	0.57	0.57
B6	Buy, sell, and trade stocks	2	0.88	0.63	2.02	0.84	0.63	2.00	0.56	0.56
B7	Decide how a company should advertise its products	2.09	0.84	0.62	2.08	0.85	0.66	2.14	0.61	0.61
B8	Sell real estate to people	1.72	0.83	0.48	1.67	0.81	0.60	2.18*	0.64	0.64
B9	Schedule meetings and appointments	2.46	0.76	0.44	2.48	0.73	0.45	1.83*	0.63	0.63
B10	Answer phone calls for a company	2.13	0.85	0.58	2.13	0.84	0.62	1.44	0.39	0.39
V1	Develop the landscaping for a home	2.2	0.86	0.54	2.23	0.83	0.50	2.12	0.52	0.52
V2	Operate heavy machinery such as a forklift or a bulldozer	1.75	0.86	0.76	1.84	0.87	0.72	1.83	0.68	0.68
V3	Install cables for electrical equipment	1.71	0.85	0.69	1.83*	0.87	0.73	1.66	0.76	0.76
V4	Fix a broken appliance with tools	2.02	0.89	0.76	2.05	0.86	0.76	2.04	0.83	0.83

Table A.1—Continued

V5	Install heating and air-conditioning equipment	2.33	0.82	0.47	2.39	0.77	0.46	1.48*	0.68	0.68
V6	Build furniture, cabinets, or other structures out of wood	1.75	0.85	0.74	1.81	0.87	0.73	2.11*	0.69	0.69
V7	Plant, grow, and harvest crops	1.74	0.84	0.75	1.88*	0.86	0.76	1.72	0.40	0.40
V8	Build and repair powerlines or wind turbines	1.9	0.87	0.73	1.96	0.86	0.75	1.51*	0.73	0.73
V9	Fix a broken car engine	1.64	0.81	0.74	1.75*	0.85	0.79	1.77*	0.70	0.70
V10	Use a torch to mold structures out of metal	2.16	0.86	0.71	2.21	0.84	0.71	1.92*	0.67	0.67

Note- All item-total correlations are significant at the α=.01 level; *=significantly different from the MTurk sample item mean at Time 1 (α=.05)

APPENDIX E

Internal Consistency (Coefficient Alpha) for Time 1, Time 2, and 8th grade samples

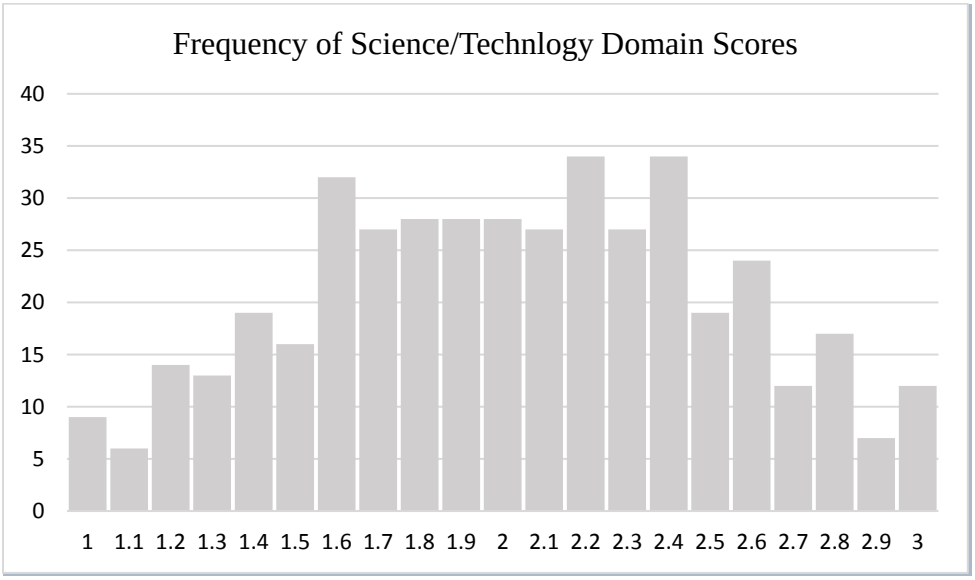
Table A.2 Internal Consistency (Coefficient Alpha) for Time 1, Time 2, and 8th grade samples

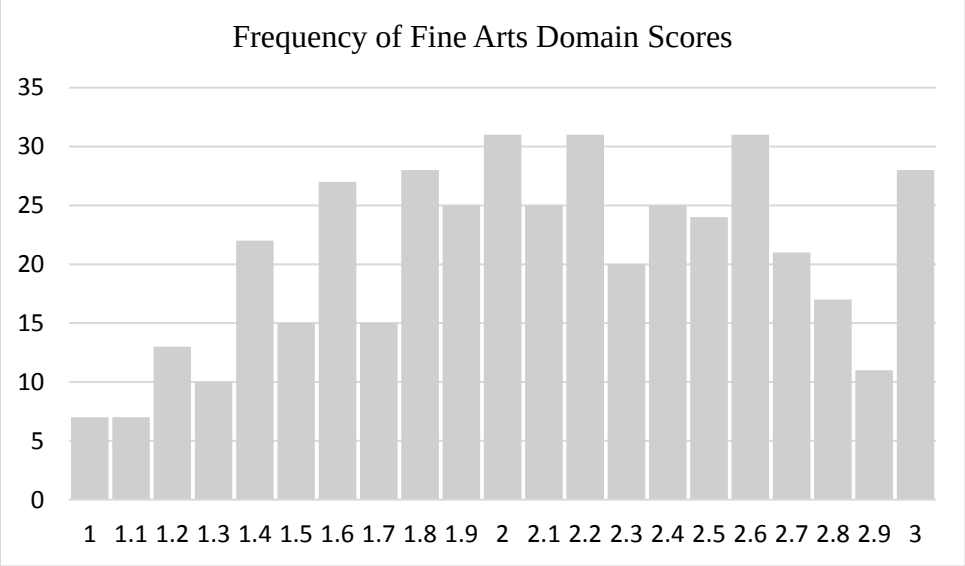
Domain	Time 1	Time 2	8th grade
Science/Technology	.77	.79	.57
Fine Arts	.81	.80	.80
Community Services	.75	.76	.70
Business	.81	.83	.80
Technical/Vocational	.88	.89	.86

Note: MTurk Time 1 and Time 2 samples only include participants who passed all indicators of attentiveness (n=426, 296, respectively). n for 8th grade sample=133.

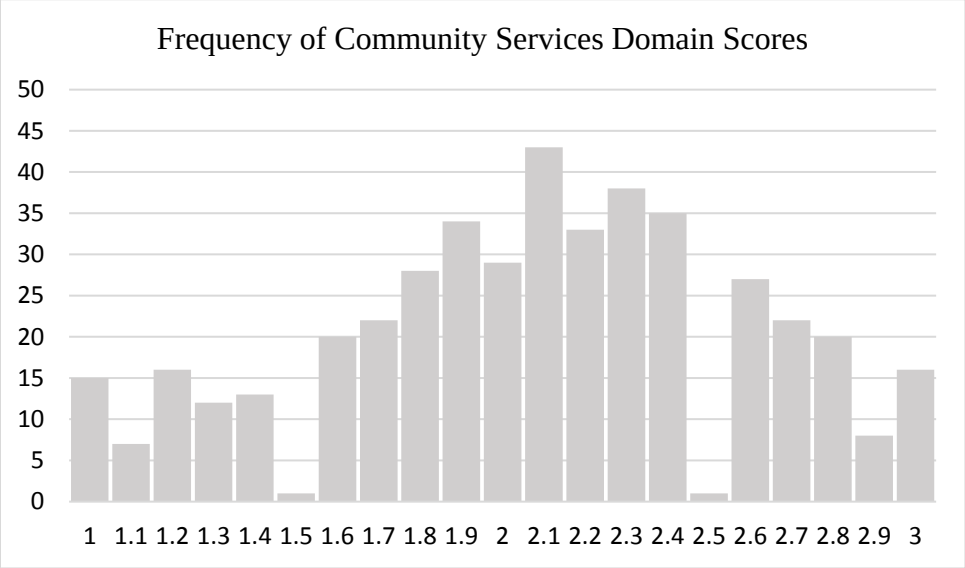
APPENDIX F

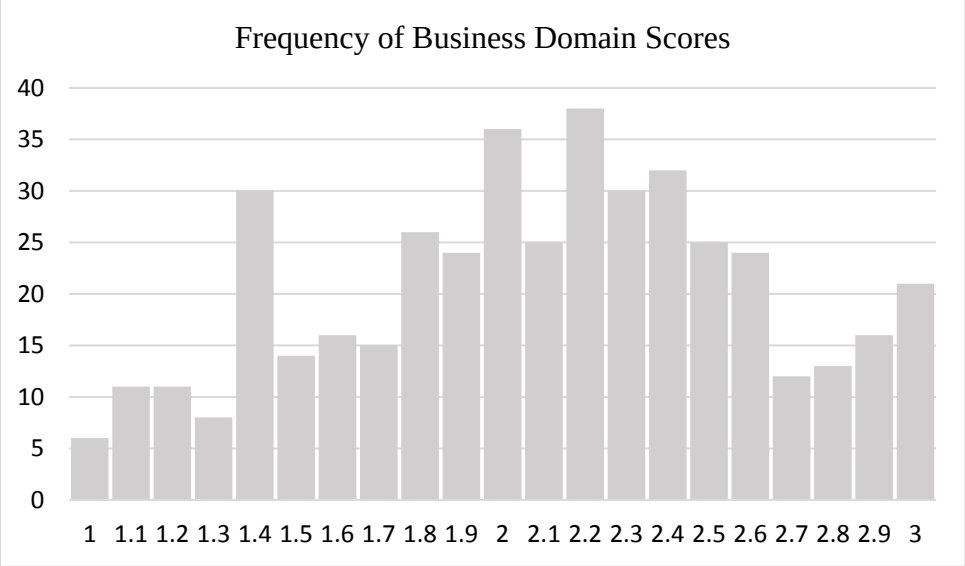
Frequency Distribution Graphs of Mean Interest Inventory Domain Scores



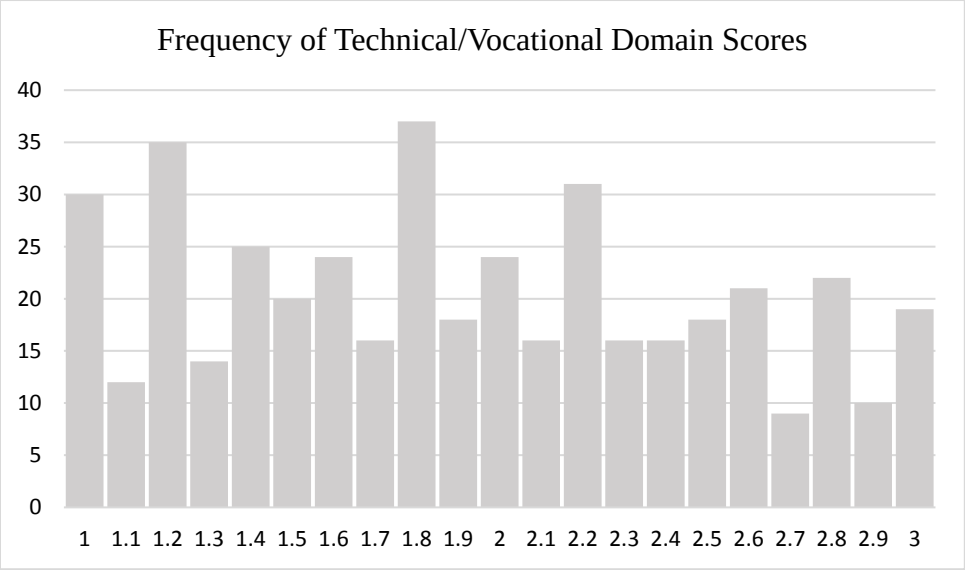


APPENDIX F (cont.)
Frequency Distribution Graphs of Mean Interest Inventory Domain Scores





APPENDIX F (cont.)
Frequency Distribution Graphs of Mean Interest Inventory Domain Scores



APPENDIX G
Coding Scheme for Match Between Highest Domain Score and Self-Reported Category of Career or College Major

After coding respondents' highest scores, a percentage of matches between those scores and self-reported category of career or college major was calculated. Participants who reported that they were unemployed or did not have a college major were excluded from this analysis. This process yielded a sample of 363 participants.

For coding purposes, a match between highest score and self-reported career or college major occurred if they were the exact same. However, some participants had two or more categories that were tied for the highest score. Ties were handled in the following manner: If an individual had two or more highest scores that were tied, and one of those scores corresponded to the same category as the self-reported career or college major, then that participant is coded as a match. If an individual had two or more scores that were tied, and neither of those highest scores corresponded with the career/major, then the individual was not coded as a match.

For example, if a participant had tied scores of 2.7 for both the Community Services and Business domains, and all other domain scores were below 2.7, and the self-reported college Major is Business, then this individual was coded as a match. If another participant had tied scores of 2.8 for Fine Arts and Technical/Vocational, and the self-reported career category was Science/Technology, then this individual does not have a match.

For college students who reported having a major, 22/52 (42.31%) had an exact match between highest interest inventory score and self-reported college major category. For non-college students who reported having a job, 103/311 (33.12%) had an exact match between highest interest inventory score and self-reported career category