How do we know someone will intervene? The validation of a survey instrument designed to measure collegiate bystander intervention disposition

Dissertation

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

By

Laura Stiltz Dahl

Graduate Program in Educational Studies

The Ohio State University

2019

Dissertation Committee

Matthew J. Mayhew, Ph.D., Advisor

Jerome D’Agostino, Ph.D.

Cecilia Mengo, Ph.D.

Abstract

Bystander intervention has emerged as a best-practice for combatting sexual violence on college and university campuses. Bystanders are those individuals who observe negative behavior and must decide whether to act in ways that benefit the perpetrator or victim, or do nothing. Although bystanders do not always act in ways which support the victim, proponents of bystander intervention education argue that equipping students with the knowledge, awareness, and skills to step in when they witness negative sexual behaviors will decrease instances of campus sexual violence as well as shift campus cultural norms. For the last decade, research and practice related to bystander intervention in collegiate contexts has been narrowly defined within the scope of sexual violence prevention, yet other types of violence are on the rise at colleges and universities across the United States. How do we know that students will intervene in these situations as well? This study attempts to address this need by examining a new instrument designed to measure bystander intervention disposition broadly across a variety of situations common to postsecondary contexts. It seeks to answer the following question: How can college student bystander intervention disposition be reliably and validly measured?

This study draws on theories of educational measurement as well as frameworks for understanding violence and violence prevention, cognitive and moral decision-making, and identity development to investigate the reliability and validity of the instrument. Bystander intervention disposition is defined as one’s innate inclination to intervene on behalf of others when faced with negative behavior and is conceptualized as a continuous latent construct.

Respondents with high bystander intervention disposition should be willing to intervene on behalf of others in situations with high costs; respondents with low bystander intervention disposition should only intervene in those situations that they find it easy to do so.

Students who responded to the 2018 administration of the Assessment of Collegiate Residential Environments and Outcomes (ACREO) were invited to read and respond to seven of 16 possible scenarios commonly found on college campuses in which negative behavior is exhibited by one or more parties. Upon reading each vignette, respondents were asked to rate their likelihood of engaging in a number of actions based on their relationship with the actors in the scenarios: knowing the victim, knowing the perpetrator, knowing other bystanders, and not knowing anyone at all. Items also spanned a variety of actions such as saying something at the time, saying something at a later time, getting others to intervene, and finding an authority figure to intervene. A total of 1,939 undergraduate students at one of three public universities responded to the items which comprise the bystander intervention disposition instrument.

Rasch analysis using Wolfe and Smith's (2007) framework was used to examine five aspects of Messick's (1995) unified concept of construct validity. Item fit; rating scale functioning and theoretical predictions; principle component analysis on the standardized residuals; person fit, differential item functioning, and person reliability; and group comparisons and person-item maps provided evidence supporting the validity of this instrument. Discussion and implications are provided.

To Stephen, D.B.M.

Thank you for your unwavering love and support as I pursue my dreams.

Acknowledgements

If I have seen further it is by standing on the shoulders of Giants. – Sir Isaac Newton

There are many people in my life who have helped make this project a reality. First and foremost, I would like to thank the wonderful members of my committee for their mentorship through this process. Dr. Jerry D'Agostino, although you joined this committee toward the end, your insight and knowledge of psychometrics and the process of validation have been invaluable. I am grateful for the conversations we have had about the Rasch model as well as starting my career as a faculty member. Dr. Cecilia Mengo, thank you for your willingness to share your content knowledge expertise and for challenging me to think more theoretically. I am a better scholar because of your encouragement and perspective. Finally, Dr. Matt Mayhew, my chair and advisor (and now friend), thank you for everything. Your mentorship these last four years has meant the world to me, and I am so thankful that we made the move to Columbus so I could continue to work with you.

This dissertation also would not have happened without the work of Zach Hooten, who helped test and administer the ACREO surveys during the Spring 2018. Zach, you are an amazing co-worker and an even better friend. Thank you for joining CoIL and for doing a fabulous job taking over ACREO from me. It is in the best hands possible, and I can't wait to see how you make it even better moving forward.

Doctoral students are lucky when they have one cohort of peers to help them through this process; I won the lottery when I got two. To my NYU cohort – Alanna, Chris, James, Maurice, Paulina, Ray, Stephanie, Susan, and Tiffani – thank you for teaching me as much as you did during that first year. I will always miss our Saturday writing sessions in the East Building conference room. A special thanks to Chris Stipeck for his wonderful friendship that continued even when I moved away. I'm not sure I would have made it without you. To my OSU cohort – Antonio, Courtney, Kaity, Lane, Shannon, and Tiffany – thank you for letting me join you with open arms. It's always difficult to transfer to a new school, and I could not have asked for a better group of people to encourage and support me during that transition and beyond. Thank you for making Columbus and Ohio State home. I am especially thankful for Antonio Duran, our social chair, who orchestrated most of our activities. You are a brilliant friend and scholar, and I look forward to working on many more projects with you. Finally, to Ashley Staples, my Ph.D. twin who was in both cohorts, thank you for making this transition with me. It would not have been the same without you. I am grateful for your always-available words of advice and your love of good beer and good coffee.

I am also grateful to those friends and colleagues who have worked with me on various research projects over the last four years. Ben Selznick, you are a never-ended source of encouragement. Thank you for showing me what is possible. I look forward to the many papers and books we will co-author together. Ethan Youngerman, thank you for entrusting me with SILLP when I first got to NYU. From our bagel and pastrami excursion in Montreal to our golf-course run in Scottsdale, I have enjoyed hanging out with you and learning from your hilarious insight. My focus on bystander intervention in higher education has a lot to do with my conversations with both of you.

I must also thank Dr. Becky Crandall, who has been a great teaching partner and a wonderful friend at Ohio State. Thank you to Drs. Zak Foste and Tiffani Polite, who helped me with way more than just navigating OSU bureaucracy. Your support and encouragement have helped me more than you will ever know. A special shout out also goes to Jenn Sheridan, who has helped me immensely with all aspects of my doctoral work. Thank you for your organization and encouragement. I'm glad you said yes to our job offer, even if it was unexpected.

To the IDEALS team at NC State, thank you for the opportunity to work with you on this awesome project. Dr. Alyssa Rockenbach, thank you for entrusting me with the dataset and analysis. It has been a pleasure to learn from you. Dr. Shauna Morin, thank you being the best postdoc/project manager for this team. We would be lost without you. Finally, to Helen, Kevin, and Lori, thank you for sharing your passion for this work. I have loved being on this team with you.

Finally, this project would not have been possible without the love, support, encouragement, and understanding of my wonderful family. Mom and Dad, thank you for instilling in me a passion for education and learning. This dissertation is the fulfillment of a collective dream. To Stephanie, thank you for sending me notes of encouragement even though you were going through a rough time yourself. We both made it through! To the Dahls – Pat, Frank, Andrew, and Rebecca – thank you for allowing me the time and space to work over the holidays. I am forever grateful for your love and understanding.

And lastly, many thanks to the love of my life, Stephen. You are the epitome of a true partner. Thank you for taking the leap and making the move to Columbus, even though it meant leaving your family and your career. Thank you for making me laugh when it got hard and taking me out for Jeni's when it got harder. You have earned every part of your doctorate by marriage. I love you.

Vita

B.S. Applied Mathematics, Georgia Institute of Technology 2009

M.Ed. College Student Affairs Administration, The University of Georgia 2011

Director of Research Programs and Advising for Undergraduate Women in STEM,

Douglass Residential College, Rutgers, The State University of New Jersey..... 2011-2015

Graduate Research Associate, Department of Educational Studies,

The Ohio State University 2015-present

Publications

Mayhew, M. J., Lo, M. A., Dahl, L. S., & Selznick, B. S. (2018). Assessing students' intention to intervene in a bystander situation. *Journal of College Student Development*, *59*(6), 762-768.

Fields of Study

Major Field: Educational Studies

Area of Specialization: Higher Education and Student Affairs

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Humans can engage in a variety of behaviors that benefit others. Some actions, such as helping someone pick up dropped books or opening a door, occur almost every day. Others, however, are extraordinary and heroic. Winners of the Carnegie Hero Fund, for instance, have been honored for acts such as jumping in front of an oncoming subway car to rescue a stranger who fell on the tracks and saving teenagers from a fiery car crash in the middle of the night (Abumrad & Krulwich, 2018). These types of behaviors, termed *prosocial,* include "the broad range of actions intended to benefit one or more people other than oneself – behaviors such as helping, comforting, sharing, and cooperation" (Batson, 1998, p. 282). Although prosocial behaviors underlie the teachings and tenants of most cultures, religions, and philosophies (Dovidio et al., 2006), the scientific study of these types of actions has only recently emerged as a recognized focus of psychology (Carlo & Randall, 2001). In fact, the term *prosocial* did not appear in most dictionaries until after social scientists coined the term as an antonym for *antisocial* (see Green, 1998). Some scholars have explored the evidence supporting a genetic basis of prosocial tendencies (see Clark & Watson, 1999), whereas others have focused on the psychology and sociology of these behaviors (see Latané & Darley, 1970). Regardless of disciplinary perspective, the study of prosocial behaviors is "important in its own right and has implications for furthering our understanding of both individual and group level processes such as morality, aggression, intimacy, interpersonal relationships, well-being, and mental pathology" (Carlo & Randall, 2001, p. 151).

When people unintentionally witness others in harmful or negative situations, they become bystanders to the incident. Victoria Banyard, one of the leading scholars of bystander intervention for preventing campus sexual violence, noted that:

> Bystanders have been defined in many different ways in both research and practice. Most definitions describe bystanders as witnesses to negative behavior (an emergency, a crime, a rule violating behavior) who by their presence have the opportunity to step in to provide help, contribute to the negative behavior or encourage it in some way, or stand by and do nothing but observe. (Banyard, 2015, p. 8)

Bystanders, therefore, are the unwitting observers to adverse circumstances who have the potential to become decision-makers and action-takers; when the bystander acts to intervene, they then engage in bystander intervention. In most cases, scholars will distinguish between bystanders who act on behalf of the victim – sometimes called “upstanders” (Ferrans et al., 2012; Twemlow & Sacco, 2013), “defenders” (Pozzoli, Gini, & Vieno, 2012), or “prosocial bystanders” (Banyard, 2011) – and those who either escalate the situation or stand by and do nothing. For the purposes of this study, however, bystander intervention is defined generally as any action taken by a third-party observer with the intention to alleviate harm.

In collegiate contexts, encouraging bystander intervention behaviors has been an emergent and promising approach to addressing the ubiquitous issue of campus sexual violence for several years (McMahon, 2010). Proponents of bystander intervention education argue that equipping students with the knowledge, awareness, and skills to step in when they witness negative sexual behaviors will decrease instances of campus sexual violence as well as shift campus cultural norms (Banyard, 2015; Hong, 2017; Korman & Greenstein, 2017; Korman, Greenstein, Wesaw, & Hopp, 2017). As such, campus administrators not only teach students how to recognize and disrupt sexually violent situations (or those with sexual violence potential), but they also emphasize the role that all community members can play in changing the campus culture to prevent sexual violence (Banyard, Plante, & Moynihan, 2004; Dovidio et al., 2006).

Since sexual violence is the most common form of violence experienced by students on college and university campuses (US Department of Education [US DOE], 2018), bystanders and bystander intervention has been studied almost exclusively as a response to this type of violence (see Banyard, 2015; Banyard et al., 2004). Campus violence, however, is not limited to sexual violence. Unfortunately, students also experience and witness other identity-based violent situations motivated by race or ethnicity, religion, sexual orientation, gender identity, national origin, and disability (US DOE, 2018) in addition to other forms of crime (e.g., theft, arson, etc.; see Carr, 2005). Since identity-based violent events are on the rise – up 25% from 2017 to 2018 (see US DOE, 2018) – students should be similarly trained to intervene to stop this form of violence as well. However, little empirical work has been done to assess student attitudes, efficacy, intervention behaviors, and willingness to intervene in these other types of situations.

Similarly, the current bystander intervention scholarship has focused on bystander intervention as a result of participation in a sexual violence prevention training. This perspective has influenced how bystander intervention is measured and assessed, with instruments narrowly focused on the topics covered in educational programs in order to determine their efficacy. Additionally, since these instruments were initially designed for assessment and not research purposes, their psychometric properties were not rigorously tested. As such, the current bystander intervention instruments are limited in their ability to explain bystander intervention beyond the scope for which they were intended. In other words, they are constrained to measuring bystander intervention educational practices in sexual violence contexts. However, scholars still use these instruments to study the relationship between students' personal characteristics and activities and bystander attitudes, efficacy, and behaviors related to sexual violence (see Bennett, Banyard, & Edwards, 2017; Foubert & Bridges, 2017a, 2017b; Hoxmeier, Adcock, & Flay, 2017)

This study attempts to move the conversation of bystander intervention in collegiate contexts away from this narrow approach and focus on sexual violence prevention. It conceptualizes bystander intervention as a psychological construct – a *disposition* – at the nexus of socio-ecological, cognitive, and psychosocial theories. Disposition was selected to describe this psychological phenomenon as it reflects one's quality of character, state of readiness, and/or tendency to act in a specified way that can also be learned. Bourdieu (1990) defined dispositions as "an acquired system of generative schemes…[that] makes possible the…production of…thoughts, perceptions, expressions, and actions" (p. 55). It is with these schemes that individuals comprehend a specific situation, determine which contextual elements are meaningful, and pursue an appropriate course of action, typically without much reflection or calculation (Weininger, 2002). In other words, bystander intervention disposition can be thought of as a continuous, latent construct that informs a person's actions in various situations, much like moral reasoning. This interpretation allows for bystander intervention to be considered across a variety of violent situations found in collegiate contexts, not only sexual violence. However, since the current bystander intervention instruments do not reflect this perspective and have weak psychometric properties, they are ill equipped to measure bystander intervention in this way. Therefore, the purpose of this study is to examine a new, innovative instrument designed to measure bystander intervention disposition as a latent construct for undergraduate students in college and university settings. Specifically, this research seeks to answer the following question: How can college student bystander intervention disposition be reliability and validly measured?

Scope of the Problem: Campus Violence

Within collegiate contexts specifically, most instances of violence on college campuses reflect the violence found in society generally (Carr, 2005). However, developing a conceptualization of campus violence is unfortunately not quite as simple as "violent acts which occur on college and university campuses" since institutions of higher education differ from other environments in which violence occurs (Roark, 1994). In fact, the expectations, ambitions, principles, missions, and values of postsecondary institutions as places committed to education and development are what make violence most out of place on campuses (Roark, 1994). The personal safety of students, faculty, and staff as they carry out their daily work both inside and outside the classroom "must be preserved if the mission of the university is to be pursued" (NASPA, 1989, p. 2).

Despite this call to keep colleges and universities safe for all members of the community, violence has unsurprisingly found a way to impact campus constituents. The pervasiveness of campus violence is the topic of many scholarly articles and books, yet little research has focused on understanding and defining this phenomenon specifically from the perspective of campus stakeholders (Mayhew et al., 2011). In an effort to develop a common understanding of campus violence, which could then equip institutional students, faculty, and staff with better tools to combat its existence and effects, Mayhew et al. (2011) explored how members of one institution understood campus violence. Their findings – which consisted of two primary themes describing how violence is more than purely physical and campus is a contextual clue – led to a comprehensive definition of the essence of campus violence:

> Campus violence is any action, verbal or physical, that coerces for the sake of harming or harms any person associated with the given campus community. It can be physical or verbal. It not only harms but coerces, often through silencing or disempowering, individuals or groups for the sake of inducing harm. It involves and affects all parts of a campus community, including the violence narrative idiosyncratic to a particular institution, its physical campus parameters, and its constituents, broadly defined as those with any stake in the given campus community. (Mayhew et al., 2011, p. 264)

Since this interpretation incorporates multiple forms of violence as well as recognizes the human aggregate aspect of campus, I prefer this conceptualization of campus violence.

One of the main issues surrounding campus violence particularly is underreporting (Carr, 2005; Mayhew et al., 2011; Sloan, Fisher, & Cullen, 1997; US DOE, 2006). Although the official rates of campus violence decreased 54% from 1995 to 2002 and students experienced most crimes at a lower rate, on average, than non-students of the same age group (Baum & Klaus, 2005), research conducted by Sloan et al. (1997) and the US DOE (2006) found that students report only 25% of violent incidents to authorities (Baum and Klaus (2005) estimated this value to be 35% for the 1995-2002 reporting period). This low level of reporting is due to a number of reasons, such as confusion over whether a situation the student faced was violent, feeling as though the violence a student experienced was not serious enough to report, and believing the violent encounter is a private matter (Pezza & Bellotti, 1995; Sloan et al., 1997; US DOE, 2006). For instance, more than 50% of campus sexual assault survivors do not report because they believe the event was not “serious enough” (Cantor, Fisher, Chibnall, Townsend, Lee, Bruce, & Thomas, 2015). Additionally, the campus environment itself may discourage the reporting of violence through policies, climate, and culture (Carr, 2005; Mayhew et al., 2011;

Roark, 1994; Slater, 1994; US DOE, 2006; Wessler & Moss, 2001). As such, "Campus violence in both its ordinary and extraordinary forms is alternately surrounded by silence or sensationalism… Yet the silence surrounding ordinary violence is much more pervasive" (Cantalupo, 2009, p. 615).

Even when one disregards the issue of underreporting, the scope of the campus violence problem is staggering. Research suggests that one in five women; one in sixteen men; and nearly one in four undergraduate students identifying as transgender, gender non-conforming, questioning, or other experience sexual and partner violence while in college (Cantor et al., 2015; Krebs, Lindquist, Warner, Fisher, & Martin, 2007), with approximately 5-15% of college men acknowledging they forced intercourse on another student (Koss et al., 1987; Malamuth, Sockloskie, Koss, & Tanaka, 1991). A significant proportion of campus sexual assaults involve the use of alcohol, are perpetrated by someone the victim knows, and occur in social settings such as residence halls or fraternities (Abbey, Ross, McDuffie, & McAulens, 1996; Basile & Smith, 2011; Fisher, Cullen, & Turner, 1999; Messman-Moore, Coates, Gaffey, & Johnson, 2008). Additionally, scholars estimate that only 11% of campus rapes are disclosed, making it the most underreported violent crime on campuses (Rand, 2009; Kilpatrick, Resnick, Ruggiero, Conoscenti & McCauley, 2007).

Although rape/sexual assault is the only violent crime against students more likely to be committed by a person the victim knows (non-strangers committed 79% of rapes/sexual assaults against students; Baum & Klaus, 2005), sexual violence is not the only form of violence from peers that students experience. Under the Clery Act, colleges and universities reported a total of

1,300 hate crimes[1] in 2016 (US DOE, 2018), a 25% increase from the year prior. Of these hate crimes, 38.4% were motivated by race, 17.9% by religion, 17.0% by sexual orientation, 15.0% by gender or gender identity, 10.8% by ethnicity or national origin, and 0.8% by disability (US DOE, 2018). However, these statistics are nothing new. In 1991, it was estimated that 30% of racial minority students were victims of violence motivated by bigotry (Abadu, 1991), with 74% of doctoral universities reporting incidents of bigotry (Cox, 1991). Ten years later, 36% of lesbian, gay, bisexual, and transgender (LGBT) undergraduate students had experienced harassment due to their sexual or gender identity, leading 20% of them to fear for their physical safety while on campus (Rankin, 2003). More recently, 54% of Jewish college students experienced anti-Semitism in 2014 (Kosmin & Keysar, 2015), a figure that reached nearly 75% in 2015 (Saxe et al., 2015). In terms of Muslim students, figures specific to college and university campuses have been challenging to find, although the Federal Bureau of Investigation (2016) reports a 67% increase in hate crimes nationally against Muslims from 2014 to 2015.

Just as general violence has multiple types (Dahlberg & Krug, 2002), violence on college campuses can manifest in a number of ways. Carr (2005) presented a conceptualization of most campus violence categories, including sexual violence (i.e., sexual harassment, sexual assault, stalking, and dating violence); racial, ethnic, and gender-based violence and homophobic intimidation; hazing; celebratory violence; attempted or completed suicide; murder and manslaughter; aggravated assault; arson; and attacks on faculty and staff. These types of campus violence can range along the continuum of violence (Kelly, 1987; Stout & McPhail, 1998) from

[1] Hate crimes are defined as "offenses motivated by biases of race, national origin, ethnicity, religion, sexual orientation, gender, or disability" (Bauman, 2018, para. 4) and may include assault, threats, or property damage (Carr, 2005). The Center for Prevention of Hate Violence (2001) argued that hate crimes are more widespread on college campuses than reported statistics.

"emotional coercion" to "verbal altercations" to "purely physical" forms of harm (Mayhew et al., 2011, pp. 261-262). Additionally, these categories are not mutually exclusive. For instance, violence impacting lesbian and gay students can take on social/emotional, physical, and/or sexual forms based on internal and external homophobia (Slater, 1994).

Several factors explain why the "campus environment provides a culture in which violence can ferment" (Roark, 1994, p. 4). Pezza and Bellotti (1995) classified these elements into three categories: predisposing factors, enabling factors, and reinforcing factors. Predisposing factors encompass the beliefs, attitudes, values, and perceptions of interpersonal violence as well as about the group of people targeted. For example, rape myths held by college men and women contribute to the preponderance of sexual violence on college and university campuses as well as its underreporting (Buddie & Miller, 2001; Burt, 1980; Comack & Peter, 2005; Du Mont, Miller, & Myhr, 2003; Eyssel & Bohner, 2011). Additionally, perceived intergroup competition, frustration, and aggression, fueled by "high 'we' vs. 'they' identity," contributes to racial-biased violence in higher education (Berg-Cross, Starr, & Sloan, 1994). This notion may also contribute to violence against religious minorities as well. As Roark (1987) summarized, "Perceiving others as of less value may be at the root of some violence and victimization – we rarely exploit equals" (p. 368).

Pezza and Bellotti (1995) described enabling factors as "the skills, resource factors, or barriers that may foster or impede the realization of behavioral predispositions" (p. 107). On college campuses, this may look like underdeveloped students experiencing independence for the first time, which necessitates negotiation of life tasks related to cognitive development and identity integration (Kitzrow, 2003; Pezza & Bellotti, 1995; Roark, 1987). Students who have not yet reached the relativistic stages of their development (Chickering & Reisser, 1993; Perry,

1968) are more likely to revert to stereotypes of others and themselves, which can lead to quicker aggressive behavior toward others (Brown & Decoster, 1989), such as conduct related to sexual conquest and other forms of traditional masculinity in men (Barnett & DiSabato, 2000; Martin & Hummer, 1989; Walters, McKellar, Lipton, & Karme, 1981) and violence directed toward underrepresented populations of students on campus (Pezza & Bellotti, 1995). The use and abuse of alcohol is another enabling factor for campus violence (Nicholson et al., 1998; Pezza & Bellotti, 1995; Presley, Meilman, & Cashin, 1997; Roark, 1994). For instance, Nicholson et al. (1998) and Presley et al. (1997) found that alcohol use on college campuses was consistently associated with sexual and non-sexual violence, including ethnic harassment, theft involving force or threat of force, physical assault, and unwanted sexual touching or intercourse. One explanation for this result as an enabling factor comes from Pezza and Bellotti (1995), who noted, “When a student commits an act of violence against another while intoxicated, it has been socially acceptable to excuse this otherwise prohibited behavior as a stereotypical reaction to the chemical effects of the drug” (p. 113).

Finally, reinforcing factors are the beliefs, attitudes, and behaviors that encourage victimization and perpetration (Pezza & Bellotti, 1995). Examples of these types of factors on college campuses include the underreporting of violence by victims (Hanson, Turbett, & Whelehan, 1986), the societal legitimization of violence (Cuomo, 1986), and the lack of perpetrator penalization by institutions for all forms of violence (Hanson et al., 1986; Mayhew et al., 2011; Payne, 2008; Roark, 1994). For instance, Mayhew et al. (2011) wondered, “What would it mean if hate speech and verbal abuse were adjudicated as forms of campus violence?” (p. 266). Most violence instances of campus hate crimes usually escalate from “lower levels of harassment” and, if left unchallenged, “the widespread use of this language may send the message that bias is accepted within a campus community" (Carr, 2005, p. 6). Microaggressions (Wing Sue, 2017), othering (Riggins, 1997), unacknowledged privilege (McIntosh, 1990), and problematic assumptions of sameness (Braidotti, 1994) at colleges and universities each contribute to institutional cultural practices, norms, and ideologies that normalize dehumanization against underserved populations and ultimately contribute to violence toward these students (Bollinger & Mata, 2018). Institutional faculty, staff, and administrators could actively change this culture (Bollinger & Mata, 2018), but confusion over policies sometimes prevents firm action by campus officials.

Student bystanders, however, can shift these campus norms and stop more explicit forms of violence through intervening behaviors. As previously stated, training bystanders to intervene is seen as a primary method of prevention for many experts in field of sexual violence since "bystander approaches attempt to develop communal responsibility for preventing sexual violence by encouraging those who are potential witnesses to take action or intervene so they can potentially challenge cultures of violence and gender inequality" (Hong, 2017, p. 29). In other words, bystanders not only act at the moment to stop violence, but scholars presume that enough of them intervening can change social norms (Banyard, 2015). Although this perspective is not shared by everyone (see Linder & Harris, 2017; Hong, 2017), bystanders continue to play a necessary role in preventing all types of campus violence. As such, more research is needed on how college student bystanders with multiple social identities respond to all types of negative behaviors and events, from low-risk to high-risk, considering the many different ecological factors that contribute to their decisions.

Theoretical Overview

Since the focus of this study is the evaluation of a new instrument to measure bystander intervention disposition, it draws on theories of educational and psychological measurement and instrument design as well as the theoretical frameworks which inform a robust understanding of collegiate bystander intervention disposition. Measurement broadly refers to the process by which numerical values are assigned to the properties of objects in order to scale and classify them. To empirically measure educational and psychological traits (i.e., latent variables), scholars must create new instruments that have been rigorously tested for reliability and validity. The instruments should also adhere to Thurstone's (1928) principles of measurement independence: object-free instrument calibration and instrument-free object measurement. The process of testing for consistency and legitimacy depends on the context in which the instrument is designed and administered; factors such as phenomenon of interest, response scale(s) of the items used to measure, and the intended population all influence the approaches used to test the psychometric functioning of a new instrument.

When it comes to understanding why, how, and when bystanders intervene on behalf of others, most scholars who study this phenomenon ground their inquiries in one of two theoretical frameworks. First, the Social Ecological Framework, developed by Dahlberg and Krug (2002) and based on Bronfenbrenner's (1970, 1993) developmental ecology model, is used to understand why violence occurs in society and how to prevent it. This framework explains the situations in which bystanders find themselves by examining "the relationship between individual and contextual factors and considers violence as the product of multiple levels of influence on behavior" (p. 12). This socio-ecological approach addresses the interaction between an individual and the environment by considering how factors such as personal characteristics, relationships with others, community membership, and societal attributes all contribute to violent behaviors. Since this perspective acknowledges the many confounding causes to violence, it encourages a multifaceted approach to stopping violence, including bystander intervention as a primary prevention measure. It is with this model in mind that experts posit the effect of bystanders: not only do they act in the moment to stop violence, but that enough of them intervening can change the social norms that encourage violence in the first place (Banyard, 2015).

The other theoretical framework commonly cited in bystander intervention research is Latané and Darley’s (1970) Decision Model of Helping. Although this framework considers the context in which negative behavior occurs, its primary focus is on the cognitive process by which bystanders arrive at intervention behaviors. In this model, bystanders must first notice the event (step 1) before interpreting it as a situation requiring intervention (step 2). They must next decide if they are responsible for acting in the situation (step 3). If so, bystanders then decide how to act (step 4) before finally implementing the intervening behavior (step 5). At any stage in this process, they may come across barriers, or inhibiting factors, which prevent them from intervening on behalf of the victim (Burn, 2009; Dovidio et al., 2006; Latané & Darley, 1970).

Piliavin et al. (1981) further refined Latané and Darley’s approach by addressing some of its limitations by integrating a cost-reward perspective to decision-making. They asserted that a bystander “analyzes the circumstances, weights the probable costs and rewards of alternative courses of action, and then arrives at a decision that will result in the best personal outcome” (Dovidio et al., 2006, p. 85). If the bystander intervenes, their action could cost them effort and time (i.e., the interruption or postponement of something important), potential personal harm, psychological aversion (i.e., the situation involves something unpleasant), financial expenditure, or social disapproval (i.e., the situation challenges a social norm). On the other hand, the helping behavior may bring monetary compensation and social benefits such as fame, gratitude, and reciprocity. Although this improved perspective is also not without limitations (e.g., humans do not always act rationally), it does sufficiently describe the decision-making processes used by bystanders.

Hoffman (1970, 2000) provided a moral perspective to bystander decision-making. His framework describes how empathy and empathy-based moral affects influence bystander intervention behaviors in order to explain how individuals resolve conflicts between caring and justice in moral dilemmas. In this approach, the moral issues regarding these dilemmas include refraining from harming others, deciding who to help when others could potentially be neglected, and determining whether to choose justice over caring (or caring over justice). Bystanders, for example, decide to help others in distress based on their level of empathy as well as the other cognitive factors described above.

Two theoretical perspectives which have yet to be fully applied to bystander intervention are also included in this study. The first is intersectionality, which emerged as an academic effort in response to the judicial treatment of Black women and women of color (Crenshaw, 1989, 1991). Although intersectionality was first used to explain how Black women experience structural, political, and representational discrimination in ways distinct from their white and male counterparts, it has materialized as a popular perspective in research, teaching, and practice for understanding how various intersecting structures of power operate for all people and social systems, not solely race and gender (Moradi & Grzanka, 2017). As a result, intersectionality has become a field of study, an analytical strategy, and a critical praxis for challenging and transforming all structures/systems of power, privilege, and oppression (Moradi & Grzanka, 2017). Intersectionality is used as an analytical strategy in this study as a way to inform the types of violent situations that occur on college campuses due to the overlapping systems of power and privilege prevalent in U.S. higher education (Linder & Harris, 2017). It additionally provides a framework for understanding who intervenes in what situations based on these systems.

The second theoretical perspective that has yet to be applied to bystander intervention is the development of a bystander identity. The Reconceptualized Model of Multiple Dimensions of Identity (RMMDI; see Jones & Abes, 2013) merges the psycho-social theories of individual meaning making (Kegan, 1982, 1994) and college student self-authorship (Baxter Magolda, 2001, 2009) with the theory of multiple dimensions of identity (Jones & McEwen, 2000). Individuals understand their core and social identities within a larger context, which is filtered by their meaning-making capacities. College and university students with more complex meaning-making capacities (i.e., self-authored attitudes and beliefs) are less likely to allow the context to determine their sense of identity and self than students with simpler meaning-making structures. This meaning-making "filter" influences bystander intervention in two primary ways. First, as student bystanders observe a given situation and context, they use their meaning-making capacities and sense of identity to make sense of the current circumstances and what possible actions to take. Second, as students are exposed to educational training programs and events on campus, they may begin to develop a bystander identity, which will influence future actions.

Organization of the Study

In the Chapter 2, I review the theories behind educational and psychological measurement, including instrument construction and the concepts of reliability and validity. I then turn to a discussion of how bystander intervention is currently measured in collegiate contexts and the issues with instruments. I discuss the theoretical frameworks and literature related to campus violence and bystander intervention which inform this study. These frameworks include theories of socio-ecological understandings of violence and environments, bystander decision-making based on cognitive and moral reasoning, and psychosocial perspectives of identity development and meaning making to acknowledge the possibility of one developing a bystander identity. Chapter 3 is dedicated to describing the research design and analytical methods used in this study. Wolfe and Smith's (2007) conceptualization of Rasch validity evidence for Messick's (1995) unified concept of construct validity served as the framework guiding the methodology used in this study. Chapter 4 provides the results of the Rasch modeling techniques that validate the collegiate bystander intervention disposition instrument. Finally, in Chapter 5, I discuss the results of this study and offer implications for theory and practice as well as recommendations for future research.

Chapter 2: Literature Review

Collegiate bystanders are thought to play an important role in preventing campus sexual violence (Banyard, 2015), and for the past decade, bystanders and bystander intervention has become analogous with campus sexual violence in both research and practice (see Banyard, 2015; Banyard et al., 2004). Since collegiate bystanders observe other forms of violence on their campuses, including identity-based violence, they should be equally equipped to intervene in these situations. However, little is known about collegiate bystander intervention in contexts unrelated to campus sexual violence. Additionally, the singular focus of using collegiate bystanders to prevent campus sexual violence has resulted in instruments designed to assess the efficacy of sexual violence prevention programs. As such, they are inappropriate for examining bystander intervention across other types of violent situations. This study expands the conversation about collegiate bystanders by considering the other forms of violence they witness on campus.

This study additionally focuses on measuring bystander intervention disposition, which has yet to be conceptualized by the collegiate bystander intervention scholarship. Disposition is defined as an individual's acquired meaning-making structure that determines their pre-reflexive thoughts, perceptions, and behaviors (Bourdieu, 1990; Weininger, 2002). Bystander intervention disposition, therefore, is considered the system of underlying psychological schemes individuals rely on when deciding whether and how to intervene when they observe harmful situations. Those with high bystander intervention disposition will intervene in more difficult situations, while those with low bystander intervention disposition will only intervene in less challenging contexts. Given this perspective, bystander intervention disposition should be conceptualized as a continuous latent construct related to bystander intervention behaviors, similar to how psychological theorist consider prosocial tendencies (see Carlo & Randall, 2001; Dovidio, Piliavin, Schroeder, & Penner, 2006).

The purpose of this chapter is to present a literature-based rationale for examining college student bystander intervention across many situations through the development of a new instrument to measure bystander intervention disposition. First, I offer a discussion of educational measurement, which pertains to the use of "a standardized situation that provides an individual with a score" (Nunnally, 1972, p. 6). Although broad in definition, educational measurement allows teachers and educators to produce tests and scales that take the "guesswork out of many types of educational decisions" (Nunnally, 1972, p. 11). The discussion of educational measurement provided in this chapter focuses specifically on the definition of measurement and the fundamentals of scale development, including establishing validity and reliability. This material is presented to provide a basis for understanding the construct of college student bystander tendencies and to illustrate the concepts behind instrument design. In the second section, I provide a critique of the existing instruments designed to measure college student bystander intervention to support the development of a new instrument designed to measure this construct. Subsequently, I explore the theoretical frameworks used to inform the concept of bystander intervention disposition my instrument is intended to measure. As an acquired system, disposition is influenced by one's context as well as personal characteristics (Bourdieu, 1990, Bronfenbrenner, 1979, 1993). As such, these frameworks include socio-ecological perspectives of violence and intersectionality, bystander decision-making processes, and psychosocial theories of identity development. This chapter concludes with the description of bystander intervention disposition developed for this study.

An Overview of Educational Measurement and Instrument Design

Educational measurement serves one of three objectives: to predict performance in a possible real-life situation, to assess efficacy of current performance, and to gauge psychological traits or constructs (Nunnally, 1972). The prediction function provides educators with a way to assess how individuals will behave or perform in the future. For example, college-entrance examinations – such as the SAT, ACT, and GRE – are used to forecast how well students will do when they actually attend institutions of higher education. The assessment function, however, "concerns a direct measurement of the effectiveness of performance at a particular point in time" (Nunnally, 1972, p. 27). Examination of student learning throughout a course is the most commonly-used instance of educational measurement fulfilling the function of assessment. Finally, educational measurement also aims to evaluate psychological traits – also called constructs – of students. Bystander intervention tendencies, which is the focus of this study, is an example of a psychological construct.

Although "all measurement… is social measurement" (Duncan, 1984, p. 35), the way researchers approach measurement depends on their discipline. Just as measurement in physical science depends on valid and reliable instruments, so too does measurement and research in social and behavioral science more broadly (DeVellis, 2017; Nunnally, 1972; Nunnally & Bernstein, 1994). However, social science measurement and research differ from physical science measurement and research in one key aspect: the role and relationship of theory. As DeVellis (2017) noted,

> Social scientists tend to rely on numerous theoretical models that concern rather narrowly circumscribed phenomena, whereas theories in the physical sciences are fewer in number and more comprehensive in scope… Measuring elusive, intangible phenomena derived from multiple, evolving theories poses a clear challenge to social science researchers. Therefore, it is especially important to be mindful of measurement procedures and to fully recognize their strengths and shortcomings. (p. 13)

For educational researchers, theory influences both what will be measured – the phenomenon – as well as how it will be measured (Pedhazur & Schmelkin, 1991). Since psychological phenomena cannot be directly measured the way physical scientists measure tangible objects, social scientists must instead rely on theory to measure a phenomenon's properties, characteristics, features, or attributes (Kerlinger, 1973). Theory, therefore, informs the relationships associated with the properties and/or attributes of the construct(s) under investigation (Lord & Novick, 2008).

The purpose of this section is to introduce a definition of measurement as well as discuss the role of validity and reliability in the development of instruments used to measure psychological constructs. Since the phenomenon of choice for this study – college student bystander intervention tendencies – is considered a psychological trait, this discussion will focus on the third function of educational measurement. Although the three functions are not entirely unrelated (Nunnally, 1972), construct validity will be featured over other forms of validity, such as face, criterion, or content, since it is the primary form of validity for measurement of psychological traits (Nunnally & Bernstein, 1994).

Defining Measurement

"Measurement is a fundamental activity of science… The process of measurement and the broader scientific questions it serves interact with each other; the boundaries between them are often imperceptible" (DeVellis, 2017, pp. 2-3). Although some scholars believe measurement occurs directly for objects (e.g., Stevens, 1951), most prefer to assign the term *measurement* instead to an object's properties and not necessarily to the object itself (see Campbell, 1928; Duncan, 1984; Jones, 1971; Nunnally & Bernstein, 1994; Torgerson, 1958). As Torgerson (1958) explained, "Properties are essentially the observable aspects or characteristics of the empirical world… Measurement is always measurement of a property and never measurement of a system[2]" (p. 9). Measurement, therefore, concerns the process by which we define the properties of systems and not the systems themselves.

This perspective of measurement holds true for all forms of science, including the physical sciences as well as the social and behavioral sciences. Constructs such as weight, length, and temperature are properties of an object. For instance, when one visits the doctor, the nature of the person is never measured. It is one's height, weight, body temperature, and blood pressure that are assessed. The same holds for scientists wanting to measure intangible phenomena such as psychological traits. We must instead focus on measuring the properties of human cognition and affect instead of attempting to measure the person himself/herself.

Measurement broadly consists of using mathematical symbols to both scale (i.e., represent the quantity of an attribute) and classify (i.e., organize objects based on a particular

[2] Torgerson (1958) defined systems as the something of which properties describe: "Thus, properties, where they occur, occur as aspects or characteristics of systems. To make the circle complete, we might define a particular system as roughly that which possesses such and so properties" (p. 9).

attribute) objects based on their properties (Nunnally & Bernstein, 1994). Although classification does not necessarily require mathematical notions (Torgerson, 1958), using quantitative concepts to measure properties has a number of advantages (Hempel, 1952). For instance, they provide greater descriptive freedom and discerning (i.e., we might classify objects as long or short, but some may be longer or shorter than others) as well as an ability to order objects based on the quantity of a property (i.e., an object that is three feet in length is longer than an object that is three inches in length). Quantitative concepts in measurement also equip scientists with more freedom to formulate general laws (i.e., as the length of an object changes, so too can other properties such as surface area and volume, *ceteris paribus*) and extensive application of higher mathematical theories and concepts (i.e., relationships between properties such as length and surface area can be stated precisely using mathematical terms). As Torgerson (1958) concluded, "the primary purpose of measurement in science is to enable us to express the functional relations between constructs in terms of mathematical equations" (p. 12).

Since measurement begins with the assignment of numerical values – most commonly those from the set of integers or real numbers (see Herstein, 1999) – to understand the properties of an object, it is important to understand the basic features and characteristics of these numbers distinct from the operations typically performed on them (Torgerson, 1958). First, real numbers are ordered. For example, the value of three is more than the value of two, which is more than the value of one. The second feature of real numbers is that the differences between them are also ordered. In other words, "the difference between any pair of numbers is greater than, equal to, or less than the difference between any other pair of numbers" (Torgerson, 1958, p. 15). Finally, the sets of integers and real numbers have an origin, usually denoted by zero. This origin also acts as a unit element (Herstein, 1999) in that it leaves other values unchanged when combined with them. As integer and real number values are assigned to the properties of objects, the relations between these values reflect the relationship between objects based on the properties. This process is how one establishes a scale of measurement (i.e., measure the property; Torgerson, 1958). A more detailed discussion of scales and scaling occurs in the next section.

In his discussion of the nature of measurement, Torgerson (1958) additionally noted the importance of distinguishing between the three kinds of information numbers represent in measurement. In other words, the kind of measurement one has depends on the ways in which numerical values obtain meaning. In the first kind of measurement, the values obtain meaning through "laws relating the property to other properties" (p. 21). Surface area is an example of this "derived measurement" (Campbell, 1920, p. 276) since it depends on the relationship between other properties (i.e., an object's length and width). The second way the characteristics of numbers obtain meaning in measurement is by definition (Campbell, 1920). Measurement of psychological traits fall into this category because we presume a relationship between the observation and the concept. The third and final kind of measurement – "fundamental measurement" (Campbell, 1920, p. 277) – relies on defined principles which relate the various quantities of a construct to each other: "A construct measured fundamentally possesses both operational and constitutive meaning of and by itself" (Torgerson, 1958, p. 22). Measurement of length and width fall into this category since these properties are of and by themselves; they need not rely on other properties to be defined.

One important aspect to consider when defining measurement is the notion of objectivity; measurement should be independent of the object being measured as well as the tool used to do the measuring. As Thurstone (1928) noted,

> The scale must transcend the group measured… One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (p. 547)

Measurement objectivity is important because it allows scientists to "generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements" (Wright & Stone, 1979, p. xii).

Thurstone (1928) described two conditions under which measurement achieves independence: object-free instrument calibration and instrument-free object measurement. Object-free instrument calibration occurs when the device used to measure an object's properties can be calibrated independently of the object. In other words, the adjustment of the measurement instrument should not depend on entity we intend to measure. For instance, although the terms *foot* and *feet* currently describe units of distance in the imperial system of measuring, they do not actually indicate that we measure people's height based on the length of their own feet. Object-free instrument calibration for measuring psychological traits allows social and behavioral scientists to construct tests with uniform meaning regardless of whom takes the test (Wright & Stone, 1979).

Instrument-free objective measurement is achieved, however, when the measurement of the object is independent of the instrument used to measure. In the physical sciences, measurement of an object's attributes does not depend on the scale used; the scale may have different units, but the ruler used does not change the length of the object. The same should occur for measurement of psychological traits:

> When we expose persons to a selection of test items in order to measure their ability, it should not matter which selection of items we use or which items they complete. We should be able to compare persons, to arrive at statistically equivalent measurements of ability, whatever selection of items happens to have been used – even when they have been measured with entirely different tests. (Wright & Stone, 1979, p. xii)

When measuring psychological constructs, scientists should aim to create instruments which do not change a person's score when the items change.

Measurement of Constructs

Measurement in the social and behavioral sciences, including education, also concerns "the process and rationale involved in the construction of a scale or measuring device and the properties that can be ascribed to it" (Torgerson, 1958, p. 13). For the physical sciences, many scales have already been established for centuries; how one measures attributes of concrete objects, such as length, time, mass, and temperature, are well known and usually undisputed. Yet, how we measure psychological constructs – traits which are themselves intangible – requires scientists to create new instruments with suitable, objective scales (Wright & Stone, 1979).

Instruments used to measure psychological traits usually consist of "items combined into a composite score and intended to reveal levels of theoretical variables not readily observable by direct means" (DeVellis, 2017, p. 15). Underpinning this approach, however, is the belief these variables exist and can be measured, although we cannot assess them directly. Defining these "latent variables" and how scientists approach "turning observations of test performance into measures of mental ability" (Wright & Stone, 1979, p. 1) is the focus of this section.

Understanding latent variables. Latent variables often refer to the underlying psychological phenomenon or construct of interest that a set of items should reflect (DeVellis, 2017). Latent variables are the opposite from manifest variables in that they are not directly observable. DeVellis additionally noted that using the term *variable* to describe these constructs recognizes them as non-constant with variations over time, place, people, or some combination of these or other dimensions. It is also important to acknowledge that latent variables are "a characteristic of the individual who is the source of the data" (DeVellis, 2017, p. 24). Thus, scientists should take care to collect data from respondents directly, either by observation or self-report, instead of relying on some form of proxy information.

Wright and Stone (1979) provide a helpful visualization of latent variables as a horizontal line with directionality indicating high ability to the right and low ability to the left. They explain that when social scientists attempt to measure a person's ability or attribute, they are actually attempting to estimate the person's location on the line implied by the latent variable (i.e., the measurement instrument will point to a specific point on the line). As such, scientists must "construct a test that defines a line" as well as "also have a way to turn the person's test performance into a location on that line" (Wright & Stone, 1979, p. 1). Items used to measure a latent variable should, therefore, all point to the same construct (i.e., fall on the same line) as well as represent different levels of difficulty (i.e., some should be harder to endorse than others). Once the items are confirmed to fit together to measure one latent variable, they can then be used to determine a person's level of the variable by locating their position on the line.

As scientists develop instruments to measure latent variables, they assume a relationship between the latent variable and the items used to gauge it; that is, the latent variable is thought to cause the item score (DeVellis, 2017). This causal relationship also implies empirical relationships between the items used on a given instrument. Since the same underlying construct correlates with each of the items, they should all correlate with each other. However, a measurement instrument for a latent variable can only estimate the actual magnitude – the *true score* – of the construct at the time and place of measurement. In other words,

> A measure of depression often conforms to the characteristics of a scale, with the responses to individual items sharing a common cause – namely, the affective state of the respondent. Thus, how someone respondents to items such as 'I feel sad' and 'My life is joyless' probably is largely determined by that person's feelings at the time. (DeVellis, 2017, p. 17)

Thus, the correlation among the items can indicate the correlation between each item and the latent variable.

Latent variables in education usually fall into one of two categories: cognitive and affective (Hopkins, 1998). Scales of cognitive latent variables focus on assessing optimum performance (e.g., how much can a student know), whereas affective scales attempt to measure a person's typical performance (e.g., how much do they usually feel). Although most educational scales focus on measurement of cognitive latent variables (Hopkins, 1998), "human feelings are important both as means and ends in education" (Tyler, 1973, p. 2). As a response to the cognitive taxonomy commonly used by educators, Krathwohl (1965) devised an affective taxonomy to organize the assessment of affective objectives. Underlying this framework is a hierarchy of internalization, with the shallowest degree of affect internalization represented by awareness and the deepest represented by characterization. Hopkins (1998) also noted that just as cognition has an affective component, so does affect have a cognitive component. However, affective questions and items differ from cognitive ones in that they do not have pre-determined correct answers; the true response to matters of personal preference are not the same for all respondents (Hopkins, 1998). As Hopkins concluded, “The correct answer to an affective question depends on the person queried; the correct answer to a cognitive question is the same for all respondents” (p. 275).

Scales and scaling. How scientists and educators measure affective latent variables largely depends on how the items are scored. Although some scholars refer to scales as the items constituting a measurement instrument (e.g., an intelligence scale; see DeVellis, 2017), this section will follow the recommendation of McDonald (1999) and define scaling as “the process of setting up the rule of correspondence between observations and the numbers assigned” with scales then acting as “the established correspondence” (p. 408). Examples of this perspective of scales and scaling include the Thurstone Attitude Scales (Thurstone, 1930) and the ever-popular Likert Scale (Likert, 1932). These scales, since they represent a correspondence of numerical values to observations, must therefore follow the axiomatic characteristics of the set of real numbers, including identity relations, order relations, concatenation, and unit relations (McDonald, 1999).

Stevens (1946) first recognized that measurement could exists in a variety of ways and, as such, measurement scales could fall into different types. The classes of measurement scales – termed *nominal, ordinal, interval,* and *ratio* by Stevens (1946) – are “determined both by the empirical operations invoked in the process of 'measuring' and by the formal (mathematical) properties of the scales" (p. 677). Understanding these operations and properties are of "great concern to several of the sciences" because "the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered" (Stevens, 1946, p. 677). In other words, the mathematical operations carried out on a measurement scale entirely depends on the specific axiomatic characteristics of numerical values by which it is described.

At the lowest level of measurement is the nominal scale. In this case, numerical values are used to label objects into categories or to identify individuals (Stevens, 1946). Although McDonald (1999) argued that this type of scale is too primitive "as not yet amounting to measurement" (p. 410), Stevens (1946) noted that "the use of numerals as names for classes is an example of the 'assignment of numerals according to rule'" (p. 679) and thus constitutes measurement as long as one does not assign the same number to different groups or different numbers to the same group. Nominal scales can be used only for classification as it requires at most the assumption of a numerical equivalence rule based on the axiomatic characteristics of identity relations, namely that either $a = b$ or $a \neq b$, if $a = b$ then $b = a$, and if $a = b$ and $b = c$ then $a = c$ (McDonald, 1999; Stevens, 1946). Thus, objects can be organized into one of a set of mutually exclusive and exhaustive categories based on whether the properties of choice are equivalent or not. For instance, if object A and object B belong in the same group based on a certain property, we could specify that A ~ B (read as "A is equivalent to B," see Herstein, 1999). As such, properties of objects would follow the above axioms: either object A and object B are in the same category (i.e., A ~ B) or not, if object B is in the same category as object A then A is in the same category as B (i.e., if A ~ B then B ~ A), and if object B is in the same category as object A and object C is in the same category as B then it follows that C is in the same category as A (i.e., if A ~ B and B ~ C then A ~ C). Given these properties, the only statistic scientists can use on nominal scales is the frequency of objects in a given category, which can then determine the most numerous class and "under certain conditions we can test, by the contingency methods, hypotheses regarding the distribution of cases among the classes" (Stevens, 1946, p. 679). We cannot, however, use the numerical labels to make any claims about the order of the categories; the numbers are only used to identify a group.

If scientists wish to rank order objects as a way to measure, they must use the ordinal scale. This scale also assigns numerical values to properties of object, but these values now "correspond to the existence of a dominance rule" (McDonald, 1999, p. 410) based on the axiomatic characteristics of order relations, specifically that if $a > b$ then $a \not< b$ and if $a > b$ and $b > c$ then $a > c$ (McDonald, 1999; Stevens, 1946). These order relations help determine if a property of an object dominates (i.e., is greater than) the property of another object by assigning numerical values with order. For instance, if object A has more of a property than object B, we could say that object $A \succ B$ (read as "A succeeds B"). It then follows that we could apply the order axioms to determine that if object A has more of a property than object B, it does not also have less of that property (i.e., if $A \succ B$ then $A \nprec B$). Additionally, if object A has more of a property than object B and object B has more of a property than object C, it follows that object A has more of a property than object C (i.e., if $A \succ B$ and if $B \succ C$ then if $A \succ C$). When it comes to statistics using ordinal scales, Stevens (1946) advised scientists to be careful with how they use the ordinal scale:

> In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of

> something more than the relative rank-order of data. On the other hand, for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results. While the outlawing of this procedure would probably serve no good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size. When only the rank-order of data is known, we should proceed cautiously with our statistics, and especially with the conclusions we draw from them. (p. 679)

This counsel subsequently signifies that social and behavioral scientists who use the Likert Scale as a measurement scale should avoid finding means and standard deviations using this data since the Likert scale does not assume equal intervals between the numerically-labeled categories.

The next level of measurement is the interval scale. Stevens (1946) noted that the interval scale uses numerical values in the traditional sense. This scale provides scientists with a way to use the additive operation on the scale since it assumes equal intervals among values. Interval scales most often also includes a zero point – although the zero point is more a matter of convenience than a true zero. Herstein (1999) would note the zero value in this case operates more as a unit element instead of a true origin. The interval scale assumes the existence of a combination rule (McDonald, 1999) based on the axiomatic characteristics of additivity found in the real number system (Herstein, 1999): $a + b = b + a$, if $a = c$ and $b = d$ then $a + b = c + d$, and $(a + b) + c = a + (b + c)$. These properties allow scientists to transform measurement values by means of addition. For instance, the quantities of a property for objects A and B can now be combined using an addition operator (i.e., A$*$B). By applying the axiomatic characteristics of additivity, it follows that the quantity of a property for object A added to the quantity of the property for object B will be the same as adding the quantity of a property for B to the quantity of the property for A (i.e., $A*B = B*A$). If objects A and C have the same quantity of a property and objects B and D have the same quantity of a property, then the concatenation of the quantity of a property for objects A and B will equal the concatenation of the quantity of the property for objects C and D (i.e., if $A \sim B$ and $B \sim D$ then $A*B = C*D$). Furthermore, the order in which we perform the operation will not change the outcome; if we add the quantity of a property for object A to the quantity of a property for object B first before then adding the quantity of a property for object C, we will get the same result as if we had added the quantity of a property for A to the previously summed quantities of the property for objects B and C (i.e., $(A*B)*C = A*(B*C)$). Most statistics can be used on interval scales, although scientists are unable to make inferences about the proportion of values (e.g., one value is twice than another) due to the lack of an absolute zero or origin (Stevens, 1946).

The final level of measurement occurs when all the other levels of measurement occur with the existence of a null element signifying an absolute zero (McDonald, 1999; Stevens, 1946). This type of scale most commonly occurs in physical sciences since the absolute zero exists as an implied value even when it may never be produced (Stevens, 1946). Use of the ratio scale requires the assumption of a null object and follows the axiomatic characteristics of unit relations (McDonald, 1999), namely $a + 0 = a$ and if $a = c$ and $b > 0$ then $a + b > c$. These properties allow scientists to transform measurement values using multiplication as well as addition since we know what means to have nothing on the scale. For instance, given the quantity of the property for an object A, there exists some zero point on the ratio scale in which the scale is anchored so that when this value is combined with the quantity of a property for object A this quantity is not changed (i.e., there exists some element I such that $A*I = A$).

Furthermore, for the quantities of a property for objects A, B, and C, if the quantity of A is equivalent to the quantity of C and the quantity of B is more than the zero point, the combined quantity of A and B is more than the quantity of B (i.e., if $A \sim C$ and $B \succ I$ then $A{*}B \succ C$). Stevens (1946) remarked that “All types of statistical measures are applicable to ratio scales, and only with these scales may we properly indulge in logarithmic transformations” (p. 680).

Understanding these levels of measurement provides the basis for designing an instrument. Since the level of measurement determines the statistics we can apply, the scale used to collect the data dictates how we analyze if the instrument measures the latent variable as well as how we use the instrument to find a person’s score on the latent variable. Thurstone (1930), for example, is credited with developing one of the first scales to measure people's attitudes toward items and situations (Hopkins, 1998). Thurstone-type attitude scales can be developed for any number of subjects. Although respondents only mark whether they agree or disagree to a series of statements, these statements have been given an intensity scale value – usually numbered from 0 to 6 (Hopkins, 1998) – that contributes to finding a respondent’s final score. The intensity scale value for each item is determined by a panel of judges who determine its level of favorability toward the subject of interest. Items are then selected for the final questionnaire given to respondents based on their level of favorability; the final set of statements should consist of items that spread evenly over the intensity scale for the subject. A respondent’s final score is found by averaging the intensity scale value across the items with which they agree. As such, the Thurstone-type attitude scaling procedure is described as the method of “equal-appearing intervals” (Hopkins, 1998, p. 276), which implies data obtained with this scale operates at the interval level.

Another scale commonly-used to measure latent variables, however, uses the ordinal level of measurement. The Likert Scale (Likert, 1932) provides respondents with a series of statements and a five-category response continuum usually ranging from strongly disagree to strongly agree; respondents are asked to select the response that best describes their reaction to each item. Since these scales are flexible and easily constructed, they are used more often than most other scales to measure affective latent variables (Hopkins, 1998). In fact, the popularity of the Likert Scale instigated a whole set of rating scales with descriptive terms specific to each particular question (e.g., "not at all confident" to "very confident" or "very unlikely" to "very likely"; see Hopkins, 1998). However, since these scales only provide respondents with options to rank order their reactions to the items, these types of scales are considered ordinal in nature; there is no guarantee that the equal distances between the response categories correspond to equal distances of the latent variable across all respondents and items (Stevens, 1946).

Instrument construction. Any instrument used to assess latent variables should fulfill the goals of measurement (i.e., object-free instrument calibration and instrument-free objective measurement) while also adhering to the purpose of fully spanning only one latent variable (Wright & Stone, 1979). As such, developing items to properly measure a latent variable requires rigorous construction and extensive, on-going validation testing. DeVellis (2017) outlined eight steps investigators should follow when developing a measurement instrument:

1. Determine clearly what it is you want to measure;
2. Generate and item pool;
3. Determine the format for measurement;
4. Have initial item pool reviewed by experts;
5. Consider inclusion of validation items;

6. Administer items to a development sample;
7. Evaluate the items; and
8. Optimize scale length.

Wright and Stone (1979), on the other hand, offered a simpler approach with only four steps:

> First, we must work out a clear idea of the variable we intend to make measures on. Second, we must construct items which are believable realizations of this idea and which can elicit signs of it in the behavior of the persons we want to measure. Third, we must demonstrate that these items when taken by suitable persons can lead to results that are consistent with our intentions. Finally, before we can use any person's score as a basis for their measure, we must determine whether or not their particular pattern of responses is, in fact, consistent with our expectations. (p. 4)

Although these two methods for developing instruments to measure latent variables seem dissimilar, they offer a common strategy: make sure you know clearly what you intend to measure before you begin, create items and responses related to this construct of various degrees of difficulty, and test the instrument for validity and reliability.

Scale development begins with a clear understanding of what the instrument should measure (DeVellis, 2017; Wright & Stone, 1979). This means starting with a clear understanding of the underlying theories relevant to the phenomenon and grounding the instrument in these perspectives (DeVellis, 2017). Investigators should also consider the level of specificity the instrument intends to measure. Specificity can vary along several dimensions related to the variable, including content domains, setting, or population (DeVellis, 2017). For instance, an instrument could be measure general levels of anxiety and depression in people, or specifically focus on assessing children’s social anxiety related to changing elementary schools midway through the academic year. Lastly, instrument developers should know how the construct they wish to measure differs from other possible constructs. This understanding ensures that the items written relate to only one latent variable (i.e., fall on a single line; Wright & Stone, 1979).

Once the phenomenon and related constructs have been determined, the next step is to create items and response options that reflect the instrument's purpose and the latent variable (DeVellis, 2017; Wright & Stone, 1979). As DeVellis (2017) noted, "Each item can be thought of as a test, in its own right, of the strength of the latent variable. Therefore, the content of each item should primarily reflect the construct of interest" (p. 109). At the beginning of this step, investigators should plan to write more items than will ultimately be used on the final instrument; the more items investigators start with, the more discerning they can be in deciding upon the final items to include in the instrument (DeVellis, 2017). One way to produce many items is by writing redundant items. Redundancy of items not only ensures the spanning of the latent variable by the items, but can also contribute to making the instrument of considerable length without adding any new information. Additionally, not all forms of redundant items are equally helpful in measuring a latent variable. Changing only the grammatical structure of an item, for instance, does not pertain to the construct and therefore does not add any useful information to the instrument (DeVellis, 2017). As investigators write items, they should only include redundant items if they believe these items will contribute to the overall measurement of the latent variable.

As researchers develop the items for an instrument, they should consider the characteristics of the items along with the response options (DeVellis, 2017). Well-crafted items generally adhere to a few best practices. First, they are not unnecessarily wordy as exceptionally lengthy items are usually more complex and unclear, which can cause confusion for respondents.

Suitable items are also written at the appropriate level of reading difficulty for the population of interest, although DeVellis (2017) recommended that most items for the general population should be written at a reading level between the fifth and seventh grades. Reading difficulty level can be assessed by counting the length of the words and syllables in a sentence as well as evaluating semantic and syntactic factors of the words used (Fry, 1977). Well-constructed items additionally avoid multiple negatives as they can add to confusion while also conveying different positions on an issue based on how they are used. These items also convey only one idea at a time (i.e., are not double-barreled), avoid ambiguous pronoun references and misplaced modifiers, and use noun and adjective forms appropriately (DeVellis, 2017).

One other consideration for constructing items is the inclusion of both positively-worded (i.e., items representing the presence of the latent variable) and negatively-worded (i.e., items representing the absence of the latent variable) items on the same instrument (DeVellis, 2017). Although inclusion of both types of items on the same instrument may avoid agreement bias in a person’s responses, changes in item polarity can actually cause confusion for many respondents. Reverse polarity of items on an instrument of a single latent variable also tend to cluster into a separate measure when empirically examined (Herche & Engelland, 1996). As such, the use of negatively-worded items with positively-worded items on a single measurement instrument should be avoided.

A final aspect of item construction needing consideration is the response format. Are the items open-ended or closed-ended? If the items are closed-ended, what particular scale will be used to capture response values? Will respondents react using checkboxes indicating their agreement as with the Thurstone (1930) scale, or will they instead rank order their agreement as with the Likert (1932) scale? An additional consideration for response options also includes the number of response categories and the concepts they indicate. For instance, the original Likert (1932) scale offers five response options corresponding to various levels of agreement from disagree strongly (valued at 1) to agree strongly (valued at 5). However, rank-order scales can have as few as three response options – or as many as 100 – and inquire about amount (e.g., none to many), confidence (e.g., very unconfident to very confident), or likelihood (e.g., very unlikely to very likely). There is also some debate among instrument designers as to the use of a neutral or middle point (e.g., "neither agree nor disagree"; see Krosnick, 1991; Sturgis, Roberts, & Smith, 2014). While there are many options from which to choose, as with the items themselves, the response options and instructions should "reflect the nature of the latent variable of interest and the intended uses of the scale" (DeVellis, 2017, p. 134).

Since "these test items become the operational definition of the variable" (Wright & Stone, 1979, p. 2), we must empirically examine whether the items validly and reliably measure the latent variable of interest before it is used in research. DeVellis (2017) suggested the items be reviewed by experts for clarity and conciseness before they are administered to a sample of respondents. Expert reviewers can also confirm the items relate to the phenomenon of interest by evaluating how relevant the items are to latent variable(s) as well as pointing out aspects of the phenomenon the items do not represent. Once the items are administered to a sample of desired respondents, the items should be evaluated for reliability, dimensionality, and performance (DeVellis, 2017; Wright & Stone, 1979). Although many statistical approaches can be used to test item reliability and validity, evaluators should choose the method of analysis based on the type of measurement scale being used and the purpose of the instrument.

Instrument reliability. Item reliability is a fundamental issue in the construction of instruments used to measure latent variables (DeVellis, 2017; Nunnally, 1972). An instrument is reliable if it “provides highly precise indications of students’ standings with respect to one another” (Nunnally, 1972, p. 79) and if it “performs in consistent, predictable ways (DeVellis, 2017, p. 39) across repeated measurements of the same person. Reliability is also considered the level of an instrument’s dependability, stability, consistency, predictability, and accuracy (Kerlinger, 1973). As Boone, Staver, and Yale (2014) noted, “‘Good’ reliability means that there is empirical evidence that an instrument, be it a survey or test, measures in the same manner from time to time (e.g., Tuesday and Wednesday), and the instrument will measure people consistently no matter their opinion (attitudes) or knowledge (test)” (p. 218). In other words, a measurement instrument with a high degree of precision and consistency is said to be highly reliable.

Determining an instrument’s reliability depends on the measurement error arising from the difference in the observed score provided by the instrument and the hypothetical true score on the latent variable (DeVellis, 2017; Nunnally, 1972). Reliable instruments should yield a score that represents some true state of the latent variable and a perfectly reliable instrument would reflect the true score and nothing else (DeVellis, 2017). However, as previously stated, measurement of a latent variable’s true score is nearly impossible; in almost all cases, the observed score obtained by the instrument will differ by some degree – the measurement error – from the true score. Instrument reliability, therefore, “is the proportion of variance attributable to the true score of the latent variable” (DeVellis, 2017, p. 39). Consequently, a score obtained on an instrument should not be considered an exact point, but rather a starting point for understanding the true score based on the instrument’s reliability (Nunnally, 1972). Thus, the more reliable an instrument, the fewer errors of measurement, and the closer respondents’ observed scores are to the value of their true scores.

Nunnally described six different possible causes of unreliability: errors due to day-to-day fluctuations, errors due to the sampling of content, errors in scoring tests, errors due to guessing, errors due to test standardization, and errors due to long-range instability. Despite the fact these sources of error relate specifically to unreliability in cognitive tests, two of them do describe why instruments of affective latent variables could be unreliable: errors due to day-to-day fluctuations and sampling of content. For instance, a respondent's mood or experiences on a particular day could influence how they respond to an instrument's items. Additionally, the number of items and the content they cover can affect an instruments unreliability. As Nunnally (1972) noted, "Unreliability inherent in most tests is due to the fact that they are not long enough and not broadly representative enough of… an aptitude trait" (pp. 101-102).

Unfortunately, measurement error is not a quantity that can be observed directly (Nunnally, 1972). As such, researchers should obtain an instrument's reliability coefficient "in such a way that it measures as many of the potential sources of error as possible" (Nunnally, 1972, p. 106). There exist several methods for estimating an instrument's reliability, including alternate-form reliability, test-retest reliability, subdivided test reliability, and internal consistency reliability (DeVellis, 2017; Nunnally, 1972). These four classes of reliability are the most common in educational settings, and as such, will be the focus of this discussion.

Alternate-form reliability. Alternate-form reliability is the most comprehensive measure of reliability (Nunnally, 1972). This approach relies on correlating respondents' scores on two alternate – yet parallel – forms of the same instrument; the stronger the correlation, the more reliable the instruments (DeVellis, 2017; Nunnally, 1972). Nunnally (1972) advised that the two instruments be administered to the same set of respondents within a 2-week period to control for the day-to-day fluctuations which can influence measurement error. Using alternate forms to test reliability also provides a good check on errors due to the sampling content; if two content samples are used on the two instruments and respondents score similarly, then it is an indication that both instruments involve relatively little sampling content error (Nunnally, 1972). Despite these benefits, alternate-form reliability can be challenging to determine because it is often difficult and time-consuming to produce and administer two forms of an effective instrument.

Test-retest reliability. If an alternate form of the instrument is not available, the same instrument can be provided twice to the same group of respondents. The correlation between their scores on the repeated administrations is the measure of test-retest reliability (Nunnally, 1972), or temporal stability (DeVellis, 2017). The rationale underlying this type of reliability estimate posits that a measure which accurately reflects the true score of a latent variable should do so on separate occasions. As such, the latent variable will influence the observed score on multiple administrations of the instrument similarly, whereas the error component will vary, allowing for the correlation of these scores to represent the extent to which the latent variable influences the observed scores (DeVellis, 2017).

Although this method requires less time and effort than constructing alternate forms, it does have a few disadvantages. First, the reliability coefficient produced will not reflect any error due to sampling content. Second, respondents could very likely remember the items from the first administration of the instrument during the second administration, which can inflate the correlation coefficient (Nunnally, 1972; Yu, 2005). As Nunnally (1972) noted, “If the retest measure of reliability is found to be .90, it would ordinarily be the case that an alternate-form measure would be less, say, .85 or .80” (p. 108). Additionally, the presence or absence of different scores across instrument administrations could be due to a variety of factors unrelated to the instrument itself (Kelly & McGrath, 1988; Nunnally & Bernstein, 1994; Yu, 2005). For instance, respondents could have experienced real change in their true score on the latent variable in between administrations, or possibly they responded to the instrument’s items in a different environment from the first to the second administration. Both of these situations could present the instrument as unreliable when in fact they are not related to the instrument’s capacity to accurately measure the true score on the latent variable (Kelly & McGrath, 1988; Yu, 2005). It is with these critiques in mind that DeVellis (2017) suggested we reconsider test-retest reliability:

> Thus, test-retest reliability, although important, may best be thought of as revealing something about the nature of a phenomenon *and* its measurement, not the latter alone. Referring to invariance in scores over time as *temporal stability* is preferable because it does not suggest, as does test-retest reliability, that measurement error is the only source of any instability we observe. (p. 69)

Split-half reliability. In the case that two administrations of an instrument, whether the same or alternate, is not feasible, researchers can instead correlate scores on two mutually exclusive parts of a single instrument administered on one occasion to obtain split-half reliability (DeVellis, 2017; Nunnally, 1972). This compromise procedure can be thought of as an approximation to finding alternate-forms reliability since the two halves of the same instrument resemble two alternate forms. There exists a variety of ways in which the instrument can be halved, including a first-half, last-half split and an odd-numbered, even-numbered split. DeVellis (2017), however, advocated for methods that ensure the halves are either balanced in terms of item difficulty or selected entirely at random. DeVellis also reminded researchers that when calculating split-half reliability, the correlation between the two halves yields an estimate for each half and not the whole instrument, and, as such, the reliability for the entire set of items is underestimated. Traditional correlation methods should instead be replaced by the Spearman-Brown approach.

Internal consistency reliability. The final class of reliability considered for this discussion is concerned with the homogeneity of the items comprising an instrument (DeVellis, 2017). Since all the items on a single instrument should measure only one latent variable, researchers should observe a relationship between the items based on their connection with the latent variable. Items that are highly intercorrelated are said to be internally consistent, and, therefore, internally reliable. As such, high inter-item correlations suggest they measure the same construct as well as form a unidimensional scale.

The most commonly-used coefficient for determining internal consistency reliability is Cronbach’s (1951) alpha (Carmines & Zeller, 1979). Computation of this value considers the proportion of the instrument’s total variance attributable to a common source (i.e., the true score on the latent variable) by comparing the variance in the total score on the instrument to the variances of the individual items. Since this approach is easy to compute and requires only a single administration of the instrument, it has become ubiquitous among researchers as the primary criterion for instrument reliability (Zumbo & Rupp, 2004). However, Cronbach’s alpha has several limitations which have prompted much critique in the last decade (see Gadermann, Guhn, & Zumbo, 2012; Sijtsma, 2009; Zumbo, Gadermann, & Zeisser, 2007).

The most common critique of Cronbach’s alpha as an indicator of reliability focuses on its use with non-continuous (i.e., non-interval level) data (Gadermann et al., 2012; Zumbo et al., 2007). Since this value is based on the Pearson correlation matrix, it should only be used with interval or ratio level data, yet rating scales, which are commonly utilized to measure continuous latent variables, constitute ordinal data (Zumbo et al., 2007). Research has shown that calculating

Cronbach’s alpha for items using Likert-type response scales with less than seven points falsely deflates the reliability estimate (Gelin, Beasley, & Zumbo, 2003; Maydeu-Olivares, Coffman, & Hartmann, 2007). As such, ordinal versions of this coefficient should be calculated using a polychoric correlation matrix instead of the Pearson correlation matrix when rating scales are used as response options to items (Gadermann et al., 2012; Zumbo et al., 2007). It should also be noted that alternative coefficients designed to estimate reliability – including McDonald’s (1985) omega, Revelle's (1979) beta, and Armor’s (1974) theta – also assume continuous data (see Zinbarg, Revelle, Yovel, & Li, 2005); the rationale for calculating ordinal versions of these coefficients is equally valid (Gadermann et al., 2012).

Instrument validity. If reliability refers to the degree of precision and consistency of an instrument, validity concerns the degree to which an instrument measures what it is designed to measure (Carmines & Zeller, 1979; DeVellis, 2017; Messick, 1995; Nunnally, 1972). An instrument is thought to be valid if it is congruent with the latent variable it intends to assess; in other words, “If a test serves its intended function well, it is said to be valid; if it does not, it is invalid” (Nunnally, 1972, p. 21). Since validity depends on an instrument’s “intended function,” it is not reported in a general sense, as in having an overall “good/bad” or “high/low” validity. Instead, an instrument’s validity can only be established with respect to a specific function of a specific variable for a specific group under a specific context (DeVellis, 2017; Nunnally, 1972).

Although scientists typically refer to validity as a property of an instrument, Messick (1995) noted that validity in fact reflects the interpretation of a respondent’s score on the instrument:

> Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions,

> but also of the persons responding as well as the context of the assessment. In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question. This is the main reason that validity is an evolving property and validation a continuing process. (p. 741)

In other words, validity should be considered as the evaluation of the evidence for and the consequences of interpreting and using the score provided by an instrument. This distinction matters since performance assessment of instruments intended to measure educational and psychological constructs is a social issue as much as it is a scientific one; concepts such as reliability and validity have meaning beyond their roles as measurement principles since they ultimately inform societal decisions (Messick, 1995).

Conventionally, scientists refer to three types of validity based on an instrument's primary function: assessment, prediction, or trait measurement (Carmines & Zeller, 1979; DeVellis, 2017; Nunnally, 1972). Content validity concerns the manner in which an instrument is constructed and can assess the efficacy of a respondent's performance at a particular moment in time (DeVellis, 2017; Nunnally, 1972). It also depends on "the extent to which an empirical measurement reflects a specific domain of content" (Carmines & Zeller, 1979, p. 20). As such, "Content validity is intimately linked to the definition of the construct being examined" (DeVellis, 2017, p. 84). For instance, teachers wishing to assess their students' addition skills rely on tests and quizzes; how well these instruments measure their students' performance in all facets of a given construct – in this case addition – is content validity.

Criterion-related validity pertains to an instrument's ability to predict a future event – or criterion – based on the respondent's responses (Carmines & Zeller, 1979; DeVellis, 2017; Nunnally, 1972). In order to have criterion-related validity, an instrument should have an empirical relationship with some established standard. Criterion-related validity is oftentimes referred to as predictive validity, since it focuses on the process of predicting future behaviors or events (Nunnally, 1972), but criterion-related validity can also concern concurrent validity and postdictive validity (DeVellis, 2017). DeVellis (2017) cautioned, however, that "Criterion-related validity by any name does not necessarily imply a causal relationship among variables, even with the time order of the predictor and the criterion are unambiguous" (p. 92). The purpose of establishing criterion-related validity, therefore, is to ensure that if a measurement instrument is supposed to serve a prediction function, it actually does.

The final conventional type of validity concerns the relationship of an instrument to measures of other related constructs (Carmines & Zeller, 1979; DeVellis, 2017; Nunnally, 1972). Nunnally (1972) claimed that construct validation is most appropriate for measures of psychological traits, with measures of these latent variables gaining meaning only after having been applied in many different contexts: "In construct validation, a new measure is tested against those variables and situations where everyone will agree that relationships would be expected" (p. 32). Carmines and Zeller (1979) additionally noted that, "Fundamentally, construct validity is concerned with the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses concerning the concepts (or constructs) that are being measured" (p. 23). In other words, construct validity uses the association an instrument has with other established measures of theoretically-related constructs to demonstrate the instrument measures the latent variable it sets out to measure.

Despite the prevalent classification of validity into these three basic types, not every scholar supports this traditional conception of validity (DeVellis, 2017). Messick (1995) instead advocated an integrated approach in which the considerations of the three types of validity are unified into “a framework for the empirical testing of rational hypotheses about score meaning and theoretically relevant relationships, including those of an applied and a scientific nature” (p. 741). This newly-defined, more comprehensive version of construct validity, which includes the concepts found in content and criterion-based validity, instead reflects six aspects of validity: content relevance and representativeness; substantive theories, process models, and process engagement; scoring models as reflective of task and domain structure; generalizability and the boundaries of score meaning; and consequences as validity evidence. Messick (1995) suggested these features, as a whole, can provide a method for addressing the many interrelated questions that arise when needing to justify the use and interpretation of a respondent’s score on a measurement instrument.

Educational Measurement Summary

This section introduced an overview of the definition of measurement in addition to discussing how social and behavioral scientists approach the creation of instruments used to measure psychological and educational constructs. Measurement, broadly defined, concerns the use of numerical values to empirically scale and classify objects based on their properties. By using values found in the real number system to quantify properties, scientists can understand how objects are similar or different, have a certain order, or relate to other objects. Although the different measurement scales provide various degrees of information and usability with statistical methods, the use of numerical values assists in clarifying the properties of all objects.

Measurement of latent variables – psychological traits which cannot be directly observed – requires a specific approach that includes the development of new instruments as well as rigorous and continuous testing of their reliability and validity. Since this form of measurement must occur indirectly, scientists must rely on several methods to insure their instruments are consistent and measure what they intend to measure. The choice of which approaches to use largely depends on the context in which the instrument is designed and administered, such as the phenomenon of interest, the items' response scale(s), intended population and respondent sample, and available resources of the researcher. As Messick (1995) concluded in his discussion of validity, instrument construction and evaluation cannot rely on only one method of evidence just as it also does not require any one form as well. However, this reality has not yet been captured in the literature on college student bystander intervention. Although several instruments have been developed in an attempt to gauge bystander knowledge, attitudes, and self-efficacy, these measurement tools have focused mostly on the predominant phenomenon of campus sexual violence and harassment. They also have a number of psychometric imperfections that limit their use as robust measurements of bystander intervention as a latent variable. In the next section, I offer a critique of these instruments before discussing the frameworks which inform a new approach to measuring bystander intervention.

Current Approaches to Measuring Collegiate Bystander Intervention

Although student bystanders can witness and disrupt all kinds of violent situations on college and university campuses (Dovidio et al., 2006), sexual violence prevention experts have emphasized the role of these students as an emergent and promising approach to ending the ubiquitous problem of sexual violence on campus (see Banyard, 2015; Korman & Greenstein, 2016; Korman et al., 2017). As higher education officials began to use bystander intervention as the primary way to end campus sexual violence, the scholarship also centered this approach in this particular context (Banyard, 2008; Banyard & Moynihan, 2011; Banyard, Moynihan, Cares, & Warner, 2014). However, the empirical literature on bystander intervention as it relates to college and university campuses focuses mostly on the efficacy of educational programs (e.g., Baynard, Plante, & Moynihan, 2005; McMahon et al., 2014; McMahon et al., 2017; Hoxmeier, McMahon, & O'Connor, 2017) or the factors contributing to student bystander attitudes and behaviors (e.g., Foubert & Bridges, 2017a; Nicksa, 2014). Although these studies have transformed the ways campuses implement bystander education programs, they rely on the same quantitative instruments, or versions of them, to measure student bystander intervention in campus sexual violence contexts. However, these scales fall short in a number of areas, including lack of rigorous psychometric testing using appropriate methods. Thus, the purpose of this section is to identify and critique the existing instruments designed to measure college student bystander disposition.

Common Approaches and Instruments

In 2005, Victoria Banyard, Elizabethe Plante, and Mary Moynihan published a groundbreaking report sponsored by the United States Department of Justice on rape prevention through bystander education (Banyard, Plante, & Moynihan, 2005). Although the aim of this project was to evaluate an innovative sexual violence prevention program, their work also resulting in a new set of instruments intended to measure student bystander knowledge, attitudes, self-efficacy, and behaviors. Building on the scholarship of Grimley et al. (1994), LaPlant (2002), Lonsway and Kothari (2000), Payne, Lonsway, and Fitzgerald (1999), and Pinzone-Glover et al. (1998) – all of whom studied the efficacy of collegiate rape prevention programs – Banyard, Plante, and Moynihan changed how college student bystander intervention programs were evaluated, and thus, how scholars measured this important tool in campus sexual violence prevention.

Since their initial report 13 years ago, these instruments – referred to as "scales" by the authors – have been reassessed and revised (see Banyard, 2008; Banyard & Moynihan, 2011; Banyard et al., 2014; McMahon et al., 2014) as they continue to be used in research and evaluation. At the same time, several other instruments have persisted as common measures of student bystander disposition without much altering, including the Illinois Rape Myth Acceptance Scale (Payne, Lonsway, & Fitzgerald, 1999), the College Date Rape Attitudes Survey (Lanier & Elliott, 1997; Lanier & Green, 2006), and Slaby Bystander Efficacy Scale (Slaby, Wilson-Brewer, & DeVos, 1994). A few scholars also developed their own instruments to answer a specific research question related to bystander intervention, but these have not been used in subsequent research (see Burn, 2009).

Although bystander intervention has dominated the scholarship on college student prosocial tendencies in recent years, instruments measuring general prosocial tendencies have been developed and used in research as well (see Carlo & Randall, 2002; De Caroli & Sagone, 2013; Kou et al., 2007; Ngai & Xie, 2018; Rodrigues et al., 2017). Given that bystander intervention, as it has been studied within the context of college campus sexual violence, is one facet of general prosocial behavior (Banyard, 2008; Carlo & Randall, 2002), this review will also consider the Prosocial Tendencies Measure (PTM; Carlo & Randall, 2002) in addition to the previously-mentioned scales focused only on bystander intervention. I begin with an overview of this measure due to its general nature and chronological placement before reviewing the Illinois Rape Myth Acceptance Scale (Payne et al., 1999) and the scales initially developed by Banyard, Plante, and Moynihan (2005). Lastly, one should note that this review is not comprehensive with respect to all the available instruments measuring college student bystander intervention disposition. Due to space considerations, this review will only focus on those scales which have had the greatest impact on the research, as assessed by continued use in the literature.

Prosocial Tendencies Measure. Carlo and Randall (2002) developed the Prosocial Tendencies Measure (PTM) in response to the lack of measures available for studying prosocial behaviors, particularly for college students. At the time, existing measures were designed with other populations in mind, such as children or early adolescents; were used in a laboratory or observational setting; and assessed either "global prosocial behavior" (p. 31) or behavior in a specific situation (e.g., sexual violence). Although evidence supporting the reliability and validity for these measures existed, Carlo and Randall conceptualized prosocial behavior differently. They also understood the limitations of observational and behavioral assessments as well as situation-specific and age-specific measures. These limitations motivated them to develop a "paper-and-pencil measure of specific types of prosocial behaviors to use with late adolescents" (p. 32).

In contrast to the conventional thinking of global prosocial behavior as personal tendencies exhibited across contexts and motives, Carlo and Randall (2002) alternatively identified six types of prosocial behaviors from the literature: public, anonymous, dire, emotional, compliant, and altruistic. It was with these categories in mind that they developed the items for the PTM from previous prosocial disposition and behavior scales (Johnson et al., 1989; Rushton, Chrisjohn, & Fekken, 1981) and prosocial moral reasoning interviews with college students (Eisenberg, Carlo, Murphy, & Court, 1995). The final version of the measure consisted of 23 items broken into six subscales: public (4 items; Cronbach's α =0.78), anonymous (5 items; Cronbach's α = 0.85), dire (3 items; Cronbach's α = 0.63), emotional (4 items; Cronbach's α =

0.75), compliant (2 items; Cronbach's α = 0.80), and altruism (5 items; Cronbach's α = 0.74). Each item asked participants to use a 5-point scale to rate the extent to which the items described them (1 = "Does not describe me at all; 5 = "Describes me greatly").

To test the psychometrics of this instrument, Carlo and Randall (2002) recruited 249 college students (104 men, 145 women) enrolled in undergraduate psychology courses at a Midwestern state university and asked them to respond to the items. Methods of analysis included multivariate analysis of variance (MANOVA) to examine gender differences in the types of prosocial behaviors as well as zero-order correlations to test the interrelations among the PTM subscales; they also employed a varimax rotated principal components exploratory factor analysis to examine the factor structure of the items. Items with factor loading of at least 0.4 were considered to load on that factor, and results indicated a six-factor structure accounting for 63.38% of the variance. Additionally, the subscales of the PTM were positively and modestly interrelated, with several of them displaying differences by gender, although no differences in the composite scale emerged.

Carlo and Randall (2002) also correlated the results from the PTM with instruments of other theoretically related constructs, including measures of prosocial moral reasoning (Carlo, Eisenberg, & Knight., 1992; Eisenberg et al., 1995), global prosocial behavior (Rushton et al., 1981), empathy (Davis, 1983), social desirability (Crowne & Marlowe, 1964), ascription of responsibility (Schwartz, 1968), and social responsibility (Berkowitz & Lutterman, 1968). Although critique of these measures is beyond the scope of this paper, it must be noted that to use these measures to psychometrically test a new instrument assumes they too are psychometrically valid. As such, results indicated the subscales correlated with these other instruments as expected, providing "evidence that the structure of prosocial behaviors and pattern of relations to other theoretically relevant variables in late adolescence is differentiated" (pp. 40-41).

In a second study, Carlo and Randall (2002) administered the PTM, along with measures of helping behaviors (Swisher, Shute, & Bibeau, 1985) and altruistic behavior (Johnson et al., 1989), to 40 college students (12 men and 28 women) twice, with two weeks between sessions. This method was used to assess test-retest reliability as well as relationships with other related constructs. Results from correlation analysis provided "evidence for the short-term temporal stability of the 6 PTM subscales and showed further evidence of convergent validity with other measures of prosocial behaviors" (p. 42).

Since 2002, the PTM has been tested and used in diverse populations. Kou et al. (2007) were the first to test the psychometric properties of the Chinese version of Prosocial Tendencies Measure (PTM-C). They found adequate reliability and good validity of this measure in the context of mainland China. Ngai and Xie (2018) expanded this research by examining the PTM-C with Chinese adolescents in Hong Kong. Their results also indicate partial support for the reliability and validity of this measure. European scholars have also used versions of the PTM with adolescents in Germany (Rodrigues et al., 2017) and Italy (De Caroli & Sagone, 2013). In the German study (Rodrigues et al., 2017), results suggest that the factor structure of the German version of the PTM-R mirrors that of the English one.

Illinois Rape Myth Acceptance Scale. Along with measures of general prosocial tendencies, researchers have also developed instruments to measure students' attitudes toward sexual violence. One topic that has received much attention from the literature related to bystander intervention is rape myths. Although rape myths as a construct was introduced by scholars such as Schwendinger and Schwendinger (1974) and Brownmiller (1975), it was first defined and measured by Martha Burt in 1980. Burt (1980) described rape myths as ‘‘prejudicial, stereotyped, or false beliefs about rape, rape victims, and rapists’’ and theorized that these myths contribute to a climate which is ‘‘hostile to rape victims’’ (p. 217). Given that rape myths influence perceptions of victims and perpetrators (Buddie & Miller, 2001; Comack & Peter, 2005; Du Mont, Miller, & Myhr, 2003; Eyssel & Bohner, 2011), which is related to bystander willingness (Burn, 2009; Loewenstein & Small, 2007), many studies of bystander intervention disposition have used scales of rape myth acceptance to assess student attitudes (see Bannon, Brosi, & Foubert, 2013; Banyard, Moynihan, & Plante, 2005; McMahon, 2010).

The most commonly used instrument to measure rape myth acceptance among college students is the Illinois Rape Myth Acceptance Scale developed by Payne, Lonsway, and Fitgerald (1999). This scale builds off the scholarship completed by Burt (1980), whose instrument focuses primarily on perceptions of the victim, and Field (1978), whose research considers general attitudes toward preventing rape, and attempts to measure a reconceptualization of rape myths as “attitudes and beliefs that are generally false but are widely and persistently held, and that serve to deny and justify male sexual aggression against women” (Lonsway & Fitzgerald, 1994, p. 134). With this survey, Payne et al. (1999) also hoped to explore the underlying structure and conceptual mapping of rape myths, an area not yet investigated on a large scale.

Creation of the Illinois Rape Myth Acceptance Scale (Payne et al., 1999) started with 120 items based on a thorough review of the literature and conversations with experts in the field. These items represented six major content areas (women lie about rape, women enjoy sexual force, women elicit or are responsible for rape, men are justified in their behavior, rape is a trivial event, and rape is a deviant event), and two underlying dimensions (justification versus denial of rape and victim versus perpetrator focus). After pretesting of these items using “traditional psychometric analysis as well as non-parametric techniques, including multidimensional scaling and cluster analysis” (p. 35), the number of items was reduced to 95 and the number of myth categories increased to 19.

To test the psychometrics of this finalized instrument, Payne et al. (1999) administered the survey in two phases to undergraduate students enrolled in psychology or educational psychology courses at a large Midwestern university. During phase one, data was collected from 604 respondents; 176 students responded in phase two. The survey was administered in same-sex groups of eight or less and also included nine oppositely-worded “filler” items to discourage response sets (p. 35). These filler items concerned rape, but were not considered myths, and were interspersed throughout the 95 rape myth items. Respondents rated their level of agreement to all the items, including the fillers, using a 7-point Likert scale (1 = not at all agree; 7 = very much agree).

Payne et al. (1999) divided the data from phase one into two sets and used iterative exploratory factor analysis on one set to test the factor structure of the 95 rape myth items. This method yielded six poorly-functioning items, that were subsequently removed, and 11 theoretically meaningful and easily interpretable rape myth components that appeared as factors: “she was careless,” “she implicitly agreed,” “she deserved it,” “it wasn’t really rape,” “he didn’t mean to,” “she wanted it,” “she lied,” “rape is a trivial event,” “rape is a deviant event,” “rape is natural,” and “rape is inevitable.” After conducting cluster analysis and principal component analysis with these 89 items, Payne et al. (1999) eliminated the items related to the “rape is inevitable” and “rape is natural” components. They also consolidated the “she implicitly agreed,” “she was careless,” and “she deserved it “into one component of “she asked for it.”

Using the second set of respondents from phase one, Payne et al. (1999) constructed and evaluated three competing models of rape myth acceptance: unidimensional, multidimensional, and hierarchical. The unidimensional model (Model 1) yielded a χ^2 (189, N = 302) of 1112, GFI of 0.70, and AGFI of 0.63, indicating a poor fit of the model to the data. The multidimensional model (Model 2) fit similarly, with χ^2 (189, N = 302) = 1446, GFI = 0.59, and AGFI =0.49. Results from the hierarchical model (Model 3), however, demonstrate good fit [χ^2 (168, N = 302) = 380, GFI = 0.91, and AGFI = 0.87] and "confirm the existence of both a substantial general factor and strong specific components of rape myth acceptance" (p. 41).

Now that the structure of rape myths was understood, Payne et al. (1999) used the full phase one dataset to reduce the scale even further to 40 items based on structural integrity, clarity, content coverage, reliability, and content weighting. The Cronbach's alpha for the full scale is 0.93 and subscale alphas ranged from 0.74-0.84 with an average of 0.79. Additionally, results indicate the items fit the hierarchical model well, with χ^2 (700, N = 604) = 1311, the GFI = 0.90, and the AGFI = 0.88. Although they considered this measure "theoretically sound and statistically well-functioning" (p. 48), Payne et al. (1999) recognized its length could be a barrier to use in research and set out to also create a short form focused on general rape myth rather than specific components. The Cronbach's alpha for the short form is 0.87. The uncorrected correlation between the full 45-item scale and the 20-item short-form scale indicated that the shorter version is a more than sufficient proxy for the full scale when assessing only general rape myth acceptance [r (602) = .97, p < .001].

Bystander education assessment. As campuses implement programs aimed at encouraging bystander intervention behavior, several quantitative measures of students' attitudes toward intervening and past intervention behaviors have been developed and utilized to assess the efficacy of these initiatives. When Banyard et al. (2005) set out to assess the efficacy of their sexual violence prevention program, few studies had experimentally evaluated prevention programs with a bystander focus, which meant little empirical literature existed on assessing bystander behaviors and attitudes. As a result, their first step was to design and pilot measures intended to fill the gaps they saw evident in the literature. Six different scales were tested in their study. Although versions of all six scales have since been used by other scholars, four have found the most utility: the Bystander Knowledge Scale, the Bystander Attitudes Scale (later renamed the Bystander Intention to Help Scale), the Bystander Behaviors Scale, and the Bystander Efficacy Scale. These four scales are the focus of this review.

During the pilot phase, Banyard et al. (2005) recruited a convenience sample of undergraduate students enrolled in an introductory psychology course at the University of New Hampshire. One group of 65 participants completed the bystander knowledge items as well as standard measures of sexual violence-related attitudes and knowledge from the literature (e.g., Illinois Rape Myth Acceptance Scale). The bystander knowledge scale was modeled after Lonsway and Kothari's (2000) program evaluation methods and consisted of 10 multiple-choice and short-answer items. Examples of knowledge items include "I know I have consent to engage in sexual behavior with my partner if…" and "Over their lifetime, approximately one in ________ men will experience sexual assault" (pp. 222-223). Although students were scored "1" for a correct response and "0" for an incorrect response, they were given the option to select "I don't know" for each of the knowledge questions.

A second group of 58 participants completed the three additional instruments along with other piloted measures. To assess bystander attitudes, Banyard et al. (2005) provided a list of 38 potential bystander helping behaviors to participants and asked them to respond on a 7-point scale their likelihood of engaging in that behavior (1 = not at all likely; 7 = extremely likely). Examples of these behaviors include “Call 911 and tell the hospital my suspicions if I suspect that my friend has been drugged” and “If I hear what sounds like yelling or fighting through my dorm or apartment walls, I talk with a resident counselor or someone else who can help” (p. 229). Scores were created for this measure by summing responses across the items. Banyard et al. (2005) used this same list of items to assess bystander behaviors, except that respondents were asked to indicate which actions they had actually taken in the last 2 months (0 = no, 1 = yes).

The bystander efficacy scale, however, was modeled on recent work by LaPlant (2002) and grounded in the literature broader self-efficacy (Banyard et al., 2005). Items on this scale described 15 bystander behaviors; participants were asked to indicate their confidence level in performing each behavior with confidence measured on a 0-100 percent scale (0% = can’t do; 10% = quite uncertain; 50% = moderately certain; 100% = very certain). Examples of efficacy items include “Express my discomfort if someone says that rape victims are to blame for being raped” and “Get help if I hear of an abusive relationship in my dorm or apartment” (p. 232). Participant scores for this scale were created by calculating the mean value for the 14 items and subtracting it from 100 so the scale represented perceived ineffectiveness.

Results from the pilot study suggest these measures “showed adequate reliability, produced a range of scores among participants without noticeable problems with skewness, and correlated in expected ways with other variables, thus supporting their continued use in the final study design” (Banyard et al., 2005, p. 86). The only reported indicators of this claim include the mean, standard deviation, range, and Cronbach’s alpha. For the knowledge scale, the mean score was 5.6 with a standard deviation of 0.89 and a range of 3-7, accounting for the “I don’t know” option. The attitudes scale yielded a Cronbach's alpha of 0.93 and a mean score of 220.46 with standard deviation of 25.35 and range of 159-266. Cronbach's alpha for the behavior scale was 0.88, while the mean score was 4.79 with standard deviation 4.82 and range [0, 25]. Finally, the efficacy scale reported a Cronbach's alpha of 0.84, and a mean score of 23.77 with standard deviation 14.74 and range [0, 60].

In the final study (Banyard, 2008; Banyard, Moynihan, & Plante, 2007; Banyard et al., 2005), these measures were reevaluated using the data from 389 undergraduate students (271 women and 172 men). Since the primary focus of this study was evaluation of a sexual violence prevention program, students were randomly assigned to one of two groups. Both groups were given pretests and posttests, but the non-control group participated in the program intervention. Banyard et al. (2005) and Banyard (2008) used this data to confirm the measures' reliabilities using both Cronbach's alpha and test-retest correlations.

The knowledge scale was updated from the pilot and now include four items with multiple parts (i.e., "indicate all that are correct"). This change increased the number of possible question items to 44. Students could still indicate that they did not know the answer and were scored with "0" for an incorrect response and "1" for a correct response. For students who attempted an answer to these items, the Cronbach's alpha was 0.84 (*M* = 17.04, *SD* = 6.12, range [0, 31], # missing = 19). Since "I don't know" was an option, a separate calculation of how many items participants indicated they did not know was also done. With 20 missing data points, the final Cronbach's alpha on this measure was 0.68 (*M* = 4.74, *SD* = 2.14, range [0, 10]). The pretest to posttest correlation for the control group was 0.76 for this scale.

For the final bystander attitudes and behaviors scales, 51 items were retained from the original 58. The scale for the attitudes items was also changed from one with seven points to a 5-

point scale; the behavior items were scored the same as before (0 = no, 1 = yes). The Cronbach's alphas were 0.94 for the attitudes scale (*M* = 198.17, *SD* = 27.77, range [73, 255], number missing = 45) and 0.89 for the behavior scale (*M* = 10.02, *SD* = 6.48, range [0, 45], number missing = 32). Additionally, the attitudes scale had a pretest to posttest correlation of 0.86, whereas the correlation for the behavior scale was 0.38. As for the efficacy scale, the items remained the same and scores were calculated the same as in the pilot. None of the students had missing data on this scale and the Cronbach's alpha was 0.87 (*M* = 20.55, *SD* = 14.19, range [0, 92.86]). Bystander efficacy had a pretest to posttest correlation of 0.81.

Criterion and construct validity were also assessed using data from this sample. Banyard (2008) correlated the efficacy and attitudes measures with the behavior measure to test criterion validity; construct validity was examined by correlating the measures with the Illinois Rape Myth Acceptance Scale. All correlations assessing criterion validity were significant at the 0.001 level. Additionally, bystander efficacy, attitudes, and behavior were each significantly correlated with rape myth acceptance in the theoretically expected directions.

Since debuting in Banyard et al.'s (2005) initial report, these measures have undergone a number of psychometric tests as scholars have attempted to fine tune their reliability and validity for research. Banyard and Moynihan (2011), for example, performed a factor analysis with varimax rotation on an updated 26-item bystander behavior measure. This analysis yielded a four-factor solution explaining 52% of the variance in the outcome. One factor, "Dealing with SV and IPV specific incidents," consisted of 12 items and explained 17.18% of variance (α = 0.85). A second factor, "Party safety," contained five items and explained 12.63% of the variance (α = 0.83). A third factor, "Helping friends in distress," consisted of five items and explained 11.04% of the variance ($\alpha = 0.74$). Finally, a fourth factor, “Confronting language,” consisted of four items and explained 10.94% of the variance ($\alpha = 0.83$).

McMahon and colleagues (McMahon, Postmus, & Koenick, 2011; McMahon et al., 2014) also evaluated revised versions of the attitudes and behaviors measures. In their first attempt to modify the attitudes and behaviors scales, McMahon et al. (2011) conducted focus groups and cognitive interviews to reduce the number items from 51 to 16. An additional “Wasn’t in the situation” option was added for the behavior items. Scoring of attitudes continued as summing across the items, but behaviors were now coded as “1” for yes, “0” for not in the situation, and “-1” for no; behavior scores were then found by summing across the items. Although only descriptive statistics are reported in McMahon et al. (2011) for these revised scales, McMahon et al. (2014) used exploratory structural equation modeling with geomin (oblique) rotation and delta parametrization to test the psychometric properties and refine the scale further. After removing seven items not deemed to be conceptually strong indicators of bystander behavior and adding 11 items to have better balance of items related to low-risk, high-risk, post-assault, and proactive situations, McMahon et al. (2014) recruited 4,386 students at a large, public university in the Northeast to complete the now 20-item scales. These authors accounted for missing data by using Markov chain Monte Carlo methods to generate five complete data sets. They also randomly split their sample into two groups to reduce the likelihood of Type I errors, with one half used for the “exploratory” analysis and the other for the “confirmatory” analysis. Results indicated very good model fit for a four-factor solution retaining 11 attitudes items: χ^2 (17, N = 2,028) = 13.82, $p < 0.001$; RMSEA = 0.08; CFI = 0.99; TLI = 0.97; WRMR = 0.67. The four factors were identified as attitudes about high-risk situations ($\alpha = 0.82$), post-assault situations ($\alpha = 0.72$), post-assault reporting of perpetrators (α

= 0.82), and proactive opportunities ($\alpha = 0.86$). For the behavior scale, results also indicated very good model fit, but for a two-factor solution retaining 10 items: χ^2 (234, N = 8,921) = 3.31, $p < 0.001$; RMSEA = 0.03; CFI = 0.99; TLI = 0.99; WRMR = 3.31. The two factors were conceptualized as intervention opportunities before, during, or after an assault ($\alpha = 0.77$) and proactive opportunities (α =0.82).

Psychometric Gaps in Current Measurement Practices

Since bystander intervention education is the most common form of preventative education on college and university campuses, accurately measuring their efficacy is of the utmost importance. This starts with designing and testing psychometrically robust instruments. Despite the strengths of these scales – their contribution to literature and breadth of information gathered – there exists many areas for improvement, particularly when it comes to psychometric testing. In this section, I critique the previously-described evidence supporting the use of these instruments.

The most common analytic methods used by these scholars to assess the psychometric properties of their scales included parametric exploration of floor and ceiling effects (i.e., using means, standard deviations, and ranges), concurrent and predictive validity confirmation through Pearson correlations with other instruments, and test-retest correlations and Cronbach's alpha to quantify reliability of the measures. (The authors also employed qualitative techniques such as literature reviews, cognitive interviews, focus groups, and conversations with experts to establish content validity. However, this paper will focus specifically on quantitative methods used by these authors.) Although these practices are not inherently incorrect, the analytic procedures used to conduct these tests have flaws.

The most egregious error made by the scholars developing and validating measures of bystander intervention disposition is disregarding the ordinal nature of Likert-type response scales for the items. Likert and Likert-type scales are very flexible and more easily constructed than most other types of attitude scales (Hopkins, 1998), but using them in analysis requires special consideration (Boone et al., 2014). For instance, the numeric values assigned to the response categories only provides a rank-order of the options; in other words, "All one knows is that selection of Strongly Agree means more agreement than selection of Agree, and all one knows is that selection of Agree means more agreement than selection of Disagree." (Boone et al., 2014, p. 7). Although the numeric codes for Likert-type choices may appear interval (i.e., linear), researchers cannot actually know if the intervals between categories are equal in size nor can they be confident the categories hold the same value for all respondents. This "ordinality" of Likert-type data poses major issues to the use of parametric methods of analysis (e.g., t-tests, regressions, ANOVAs, etc.) as these tests assume the data is ratio in nature.

Since all of the authors cited above analyzed raw survey data to test the psychometric properties of their instruments, they were most likely in violation of the methods' assumptions of linearity (Boone et al., 2014). This unfortunately implies that every piece of evidence surrounding the validity of the factor score(s) generated by the instruments – from the MANOVAs testing for difference between gender-based groups to the Pearson correlations against other measures to the use of means, standard deviations, and ranges to explore floor and ceiling effects – should be treated with some level of caution. Ignoring the ordinal nature of the raw data can cause biased results for these tests. Additionally, Cronbach's alpha, the most common indicator of reliability, is only appropriate for continuous data and can underestimate the true reliability when ordinal data is instead used.

Disregarding the ordinal nature of their survey data also influences the results of factor and components analyses since these methods were designed for continuous data. Exploratory factor analysis and principal components analysis – which was employed by Carlo and Randall (2002), Payne et al. (1999), and Banyard and Moynihan (2011) – use correlation matrices to evaluate the factor structure of the items, whereas confirmatory methods and structural equation modeling – techniques used by Payne et al. (1999) and McMahon et al. (2014) – is performed on either a covariance or correlation matrix. In either case, Pearson correlations assume the raw data is continuous, or at least approximately continuous, to determine linearity. When the data is not continuous, as with Likert-type response scales, the validity of the EFA/CFA is constrained by serious limitations.

Another critique leveraged against these authors' methods concerns the use of principal components analysis and orthogonal rotations when examining the factor structure of the items. For example, Carlo and Randall (2002) examined the interrelations between their hypothesized subscales – which showed the subscales of the PTM were positively and modestly interrelated – yet conducted a varimax (i.e., orthogonal) rotation to interpret the factor structure. An oblique rotation would have been a more appropriate choice since a relationship clearly exists among the subscales. This oversight was also committed by Banyard and Moynihan (2011) as they examined the revised version of the Bystander Behavior Scale.

The disregard of other analytic methods by all the authors in some way is another issue with these measures. For instance, McMahon et al. (2014) were the only authors who conducted appropriate missing data analyses and accounted for it using multiple imputation. Not accurately addressing missing data can lead to biased results and should be avoided. Additionally, although these authors attempted to evaluate floor and ceiling effects using means, standard deviations, and ranges, none of them attempted to find gaps other places in their measurement of the latent construct or test the difficulty of their items. This lack of information, coupled with the limitations of using ordinal data to parametrically test floor and ceiling effects, leads to serious questions of the validity of the factor scores to measure the intended latent construct.

Lastly, these measures have not been tested for sample or item independence and could be subjected to bias based on the mostly small convenience samples on which they were tested. A measurement instrument can be considered objective if and only if it is independent of the object it intends to measure and if the property measured is also independent of the instrument (Thurstone, 1928; Wright & Stone, 1979). Objective instruments are important to educational and psychological research because they "make it possible to generalize measurement beyond the particular instrument used" (Wright & Stone, 1979, p. xii). More importantly, "The growth of science depends on the development of objective methods for transforming observation into measurement" (Wright & Stone, 1979, p. xi). However, no attempt was made by the creators of these bystander measures to assess objective measurement, leading us to question their validity in accurately measuring the intended construct altogether.

Bystander Measurement Summary

Designing a new psychometric measurement instrument is never an easy task, especially if it involves a topic as sensitive as bystander intervention. However, the prevalence of violence at American colleges and universities calls on scholars and practitioners to make sure they are doing everything they can to make campuses safe for all students. Educating and encouraging students to create a campus culture and community which supports victims and holds each other accountable to intervene in potentially dangerous situations is a powerful solution to ending all forms of campus violence.

As bystander intervention education continues to dominate higher education conversations, scholars and practitioners need more reliable measures of student bystander disposition. This call for more refined instruments has been echoed by several key scholars in the field (see Hoxmeier et al., 2017; McMahon et al., 2017). However, these authors generally seem satisfied with current measures of attitudes and have instead focused their efforts on improving bystander action measures. I would argue that with the current psychometric limitations of the current scales, instruments of both bystander attitudes and behaviors need to be reevaluated and quite possibly redesigned. With more information about when, how, and why students will act, higher education professionals can design more effective interventions to end campus violence, ultimately making our campuses, and subsequently our society, safer for all.

Theoretical Frameworks of Bystander Intervention Disposition

Bystander intervention emerged as a topic of scientific study after the 1964 public murder of Kitty Genovese in New York, in which numerous people were said to have witnessed the assault but not one attempted to intervene (Dovidio et al., 2006; Manning, Levine, & Collins, 2007). This incident caught the attention of social psychologists John Darley and Bibb Latané, who pioneered the study of bystander responses in emergency and non-emergency situations (see Latané & Darley, 1970). Since then, scholars have prioritized the role community members play in responding to emergencies and most forms of violence (Baynard et al., 2004; Dovidio et al., 2006).

Given the many disciplines that have explored bystander intervention and prosocial behaviors more generally (Dovidio et al., 2006), it comes as no surprise that scholars have also utilized a variety of approaches and frameworks to understand who bystanders are and how they respond. The purpose of this section is to introduce the theoretical frameworks that inform my understanding of bystander intervention disposition. Since disposition is related to context, I first describe how violence is theorized using the socio-ecological framework, the continuum of violence, and intersectionality. I then outline the theories related to bystander decision-making, including the Socioecological Developmental Model of Prosocial Action (Carlo & Randall, 2001), The Decision Model of Helping (Dovidio et al., 2006; Latané & Darley, 1970), and morality and moral development (Batson, 1998; Hoffman, 2000). I close with the components of the Reconceptualized Model of Multiple Dimensions of Identity to frame how bystanders' identity development informs their meaning-making of violent situations and intervention.

Theoretical Perspectives of Violence

Bystanders have been defined as those persons who have the potential to disrupt negative behaviors, but what does "negative behavior" mean? Is a parent responding to their child's cries a bystander? What about a person who helps another pick up accidentally dropped books? Or someone who returns a lost $100 bill to its rightful owner? Many scholars would agree that these actions represent prosocial behaviors (see Batson, 1998; Dovidio et al., 2006), but they may not amount to bystander intervention. Instead, "negative behavior" is interpreted as those actions or situations which could cause physical or emotional harm to oneself, another person, or group of people.

This definition mirrors that given to violence by the World Health Organization (1996): "The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either result in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation" (as cited in Dahlberg & Krug, 2002, p. 5). This comprehensive definition focuses three important aspects of violence: intentionality, power, and outcomes. By including intentionality in this definition, the WHO can exclude unintended incidents – such as traffic accidents – from their understanding of violence. However, intentionality also can add a level of complexity to this definition. As Dahlberg and Krug (2002) noted, “even though violence is distinguished from unintended events that result in injuries, the presence of an intent to use force does not necessarily mean that there was an intent to cause damage. Indeed, there may be a considerable disparity between intended behavior and intended consequence” (p. 5). Another level of complexity stems from the fact that intent to injure is not always the same as the intent to use violence (Dahlberg & Krug, 2002) since violence is determined differently based on cultural backgrounds and beliefs (Mead, 1969; Roark, 1994; Walters & Parke, 1964). Mead (1969), for instance, defined violence as “behavior designed to damage persons, property, or institutions which constitutes a break in expected and sanctioned behavior and is experienced by other members of the same culture as a positive violation of culturally patterned interpersonal behavioral norms” (p. 227). In other words, a person may intend to harm another, but this action may not be recognized as violent due to community cultural norms and practices. Despite this, defining violence should focus on the health and well-being of the individual victim, with the intention to harm therefore indicating an intention to be violent.

The other two important aspects of this definition of violence are use of power or physical force and the extensive range of possible outcomes (Dahlberg & Krug, 2002). By recognizing the use of power or physical force, the WHO “broadens the nature of a violent act and expands the conventional understanding of violence to include those acts that result from a power relationship, including threats and intimidation” (Dahlberg & Krug, 2002, p. 5). Since those with power can act violently by way of active perpetration as well as negligence, “the use of physical force or power” should include instances of neglect, all types of physical and psychological abuse, and self-abusive acts. Additionally, violence is not limited to actions which cause immediate injury, disability, or death; violence also includes those behaviors which result in psychological harm, deprivation, and maldevelopment (Dahlberg & Krug, 2002). Gregg (1966), for instance, posited that the definition of violence “must include not only physical acts but verbal, psychological, symbolic, and spiritual attacks, with many forms taking on a combination of characteristics” (Roark, 1994, pp. 8-9). In this sense of violence, “verbal” attacks include written and oral expressions meant to demean and humiliate others, “psychological” attacks are actions which deny someone his or her humanity and equality, “symbolic” attacks encompass behaviors which elicit fear and hostility in individuals without physical violence, and “spiritual” attacks cover demonstrations of hostility and hatred toward people, particularly those that communicate inferiority and worthlessness (Gregg, 1966).

In addition to providing a definition of violence, the WHO also developed a typology of violence based on who commits the violent act and how (Dahlberg & Krug, 2002). This categorization differentiates violence into three broadly-defined groups by who commits the violence: self-directed violence, interpersonal violence, and collective violence. Self-directed violence occurs when a person inflicts violence upon oneself and can be further subdivided into suicidal behaviors (including thoughts and attempts) and self-abuse/self-mutilation (Dahlberg & Krug, 2002). Interpersonal violence, on the other hand, is violence inflicted by an individual or small group of individuals on another person (Dahlberg & Krug, 2002). This type of violence is also divided into two sub-categories. Family and intimate partner violence usually transpires in the home between family members and intimate partners and includes child and elderly abuse in addition to domestic abuse. Community violence happens outside the home between individuals who are unrelated and may or may not know one another. Youth violence, random violence, sexual violence by strangers are all examples of community violence. The final type of violence, collective violence, occurs when violence is inflicted by larger organizations, such as governments, organized political groups, militia groups, and terrorist organizations (Dahlberg & Krug, 2002). Collective violence can be subdivided into three groups based on motive. Social violence is committed by large groups to advance a particular social agenda. Politically-motivated violence, however, includes acts of war and state violence, whereas economic violence happens when large groups are motivated by economic gain, such as disrupting economic activity or causing economic fragmentation. It should be noted, however, that in many instances, collective violence can have more than one motivation (Dahlberg & Krug, 2002).

The WHO's typology of violence also recognizes the many ways in which violence can be inflicted. Violent acts can be physical, sexual, or psychological in nature or, as previously mentioned, involve deprivation or neglect (Dahlberg & Krug, 2002). These four natures of violence can occur in conjunction with each of the three types of violence, the exception being self-directed violence, which cannot be sexual. For instance, domestic violence includes physical, sexual, and psychological abuse, as well as neglect. Sexual harassment in the workplace, physical assaults by a stranger, and neglect of elders in care facilities are all examples of interpersonal community violence. And rape during political conflicts and physical and psychological warfare can occur as collective violence (Dahlberg & Krug, 2002).

Socio-ecological framework. Several scholars have used the social-ecological framework, which draws heavily from ecological theory (Bronfenbrenner, 1979), to examine the origins of violence (see Campus Technical Assistance and Resource Project, n.d.; Dahlberg & Krug, 2002; Swearer & Espelage, 2004). These frameworks also draw heavily on human ecology theory to understand how the "interaction and interdependence of humans (as individuals, groups, and societies) with the environment" contribute to our understanding of violence and violence prevention (Bubolz & Sontag, 1993, p. 421). This theory places everyone associated with violence at the center of several nested contexts, including the human-built environment, the social-cultural environment, and the natural physical-biological environment (Bubolz & Sontag, 1993). Scholars in the applied fields of public health, social work, and education have used this theory to understand how violence and human development are "encouraged and/or inhibited as a result of the complex relationships between the individual, family, peer group, school, community, and culture" (Swearer & Espelage, 2004, p. 3).

Urie Bronfenbrenner (1979, 1993) pioneered the use of person-environment theories with his adaptation of Lewin's (1936) equation $\mathrm{B} = f(\mathrm{P}, \mathrm{E})$ to focus on human development as an outcome instead of human behavior. Bronfenbrenner rejected the idea that human characteristics, such as intelligence and moral reasoning, can be assessed objectively and without consideration for the individual's context. Instead, his developmental ecology model posits that the interaction among four primary components – process, person, context, and time – are what promote or inhibit development.

As individuals interact with their environment, they participate in the process component of Bronfenbrenner's model (1993). Process is the primary means by which development occurs, yet not every interaction will result in development. Bronfenbrenner asserts that individuals must engage in increasingly complex actions and tasks in order for development to occur, and this interaction must take place in the immediate, "face-to-face" setting surrounding the person (p. 10). Personal characteristics also shape development, with dispositions toward the immediate environment – called *developmentally instigative characteristics* – holding the most influence. Bronfenbrenner identified four different types of developmentally instigative characteristics:

those which invite or inhibit responses from the environment (i.e., elicit certain responses from others), those which define how individuals react to and explore their environment, those which relate to how individuals engage in increasingly complex activities, and those which refer to an individual’s perception of agency within an environment.

Despite the important influences of processes and personal characteristics on development, the contextual element of this framework often receives the most attention from researchers (Renn & Arnold, 2003). Bronfenbrenner (1993) described the context as a series of nested environmental systems, each inside the other with the person at the center, and which spread outward as the level of direct interaction with the person decreases. Although violence scholars focus on how these systems exert influence on the individual (see Dahlberg & Krug, 2002; Swearer & Espelege, 2004), Bronfenbrenner also acknowledged the ability of the individual to influence their context. It is within these various systems that humans, with their developmentally investigative characteristics, interact back and forth with aspects of their environment, influencing development.

At closest proximity to the individual is the *microsystem*, which is defined as “a pattern of activities, roles, and interpersonal relations experienced by the developing persons in a given face-to-face setting with particular physical, social, and symbolic features that invite, permit, or inhibit engagement in sustained, progressively more complex interaction with, and activity in, the immediate environment" (Bronfenbrenner, 1993, p. 15). In other words, the microsystem consists of one’s immediate interactions with others. Since these microsystems describe one’s closest surroundings, individuals very likely have more than one microsystem in which they operate. For example, a person’s job or educational environment may be a separate microsystem from their home or personal lives. Additionally, as individuals move between microsystems, they encounter different forces which either promote or inhibit their development.

These various microsystems interact with one another within the *mesosystem*. Renn and Arnold (2003) described the mesosystem as the context in which “the effects within and across systems may reinforce one another or they may act against one another, drawing attention to discrepancies and causing the [individual] to confront contradictory processes and messages between individual microsystems” (pp. 270-271). Beyond the mesosystem exists the exosystem, which consists of those environmental settings not part of the individual’s immediate context, yet still exerts influence on their developmental possibilities. The exosystem can include the mesosystems of others with whom the individual has contact as well as the federal and state governments, news and media, or the general economy. Lastly, the *macrosystem* describes the most distant environmental influences and "consists of the overarching pattern of micro-, meso-, and exosystems characteristic of a given culture, subculture, or other extended social structure, with particular reference to the developmentally instigative belief systems, resources, hazards, lifestyles, opportunity structures, life course options and patterns of social interchange that are embedded in such overarching systems" (Bronfenbrenner, 1993, p. 25).

The final element of Bronfenbrenner’s (1979, 1993) framework is time, which is represented by the *chronosystem*. The chronosystem accounts for how the presence and passage of time also exert influences on the individual. For example, an individual’s role within a microsystem, as well as the microsystems they inhabit, change as they age. Federal policies and cultural norms also shift with time, meaning an individual’s entire context, including micro-, meso-, exo-, and macrosystems could look completely different over the span of a lifetime. The chronosystem is also instrumental in personal development. As people interact with their environments over time, they mature and their interactions with the environment changes.

In addition to revolutionizing the way social psychologists conceptualize human development, Bronfenbrenner's theory has also informed how public health experts and social workers examine violence (see Dahlberg & Krug, 2002; Swearer & Espelage, 2004). Since this ecological model reflects the complex social exchanges between an individual and the environment, it considers how factors such as personal characteristics, relationships with others, community membership, and societal attributes contribute to violent behaviors. From this perspective, violence is acknowledged as "the product of multiple levels of influence on behavior" (Dahlberg & Krug, 2002, p. 12).

The first level of Dahlberg and Krug's (2002) ecological model for understanding violence consists of the biological and personal history factors that influence an individual's behavior. Features such as personality, educational attainment, substance abuse, and prior history with violence are all associated with one's likelihood to engage in violent behaviors. Additionally, certain types of proximal social relationships – the second level of the model – can also increase the risk for violent behaviors or victimization. The third level of the model extends the influence of relationships into the community contexts in which they are embedded. Examination of this level seeks to understand what attributes of a school, neighborhood, or workplace are associated with violence. The final level focuses on the larger societal factors which impact violent behaviors. Societal factors can consist of cultural norms which accept violence as a way to resolve conflicts, attitudes toward others which promote or condone violence, or "the health, educational, economic and social policies that maintain high levels of economic or social inequality between groups in society" (Dahlberg & Krug, 2002, p. 13).

Experts have also used this model to inform violence prevention efforts by taking a multifaceted response (Dahlberg & Krug, 2002; Campus Technical Assistance and Resource Project, n.d.). Instead of focusing on secondary prevention measures – those which occur in immediate response to violence – for victims of violence, researchers are now emphasizing the use of primary prevention measures to shift cultural norms, adjust organizational policies, and influence personal relationships to prevent violence before it occurs. The Campus Technical Assistance and Resource Project (n.d.) has found this framework particularly useful for understanding and preventing sexual violence on college and university campuses. They provided examples of actions campus community members can take at each level of the model to support a comprehensive approach to end campus sexual violence, which include: "Attend training to increase bystander intervention skills" (individual level), "Third and fourth year students get friends home safely and model active bystander behavior for incoming students" (relationships level), "Coach has zero tolerance policy among players" (organization level), and "Local bars implement training for bartenders" (community level; p. 11).

The continuum of violence. As helpful as the socio-ecological framework is for understanding the various forms violence can take by "capturing the nature of violent acts, the relevance of the setting, the relationship between the perpetrator and the victim, and – in the case of collective violence – possible motivations for the violence" (Dahlberg & Krug, 2002, p. 7), it fails to explicitly recognize how violence can range from minor/ordinary behaviors to the severe/extraordinary. (It should be noted the Dahlberg & Krug framework does implicitly account for the continuum through acknowledgement of individual behaviors on the societal factors contributing to violence and the effect of cultural norms on individual actions). One interpretation of this range of situations comes from Kelly (1987) and Stout and McPhail (1998), who specifically focused on sexual violence. Their framework, called the *continuum of sexual violence*, was “based on a feminist perspective that conceptualizes various forms of sexual violence against women not as separate, discrete acts but rather as connected and all based in patriarchal power and control” (McMahon & Banyard, 2012, p. 4). Through the use of the continuum of violence conceptual framework, scholars recognize that there exists a range of linked sexually violent behaviors that can escalate in severity (Kelly, 1987, 1989; Leidig, 1992; Osborne, 1995; Stout, 1991). Low-risk behaviors consist of actions which perpetuate myths around sexual violence or contribute to the existence of sexual violence, such as harassment, sexually degrading language, or sexually violent media images. They are considered “low-risk” due to the perceived low potential to cause harm to a victim, yet these behaviors contribute to a culture of violence, which in turn supports and condones the more severe and violent behaviors at the high-risk end of the continuum (e.g., as rape, sexual assault, and criminal sexual contact). For instance, research by Foubert, Brosi, and Bannon (2011) indicated that pornography-viewing by college men is linked to decreased likelihood to intervene as a bystander and increased behavioral intent to rape.

Although this conceptualization of violence as a continuum was developed to describe violent behaviors toward women, it can also apply to other cases of violence, including bias-related violence (Bollinger & Mata, 2018; Wessler & Moss, 2001) and campus shootings (Cantalupo, 2009). For instance, low-risk violent behaviors such as racial microaggressions or degrading language against a religious group in the media contribute to high-risk violence toward people of color, religious minorities, and members of the LGBTQ+ community (Bollinger & Mata, 2018; Wessler & Moss, 2001). Additionally, Cantalupo (2009) claimed that by examining “ordinary” (p. 613) forms of violence, scholars can understand less common, yet more extreme, violent behaviors. The notion of violence as a continuum of escalation also underlies the “broken windows theory” of crime (Wilson & Kelling, 1982); by stopping smaller crimes, law enforcement can help prevent major crimes.

Nicoletti, Spencer-Thomas, and Dvoskina (2018) additionally used the notion of a continuum to theorize how individuals verbalize and enact escalating forms of violence. The “verbal abuse continuum” begins with compliant verbal statements – those which reflect normal verbal interactions through cooperation and compliance (Nicoletti et al., 2018, p. 50). Along the continuum, statements can escalate to negative (i.e., generally pessimistic with frequent complaints) to abusive (i.e., generally disrespectful and blameful of others) to derogatory (i.e., offensive and “marked by vulgar, racist, sexist, and slanderous words” intended to objectify others; Nicoletti et al., 2018, p. 51) statements. The final stage of this continuum is verbally assaulting/threatening statements, which are clearly hostile and intending to harm or intimidate. Nicoletti et al. (2018) noted the importance of recognizing and responding to statements at all level of the verbal abuse continuum:

> The lower levels of verbal abuse are designed to manipulate, intimidate, and otherwise control the behavior of others. These statements suggest the individual has minimal coping and/or interpersonal skills. While the lower levels of verbal abuse are not particularly dangerous, they do require attention and monitoring. The individual who makes verbally abusive, derogatory, or assaulting statements is a serious threat to the campus. This individual is likely experiencing an intense level of rage which could result in impulsive or destructive actions if timely intervention does not occur. (p. 51)

In addition to the verbal abuse continuum, Nicoletti et al. (2018) articulated a physical abuse continuum, which ranges from compliant to deadly assault. Compliant behavior reflects normal physical conduct and is the least threatening form of physical harm. As individuals intensify in physical abuse, they exhibit passive resistant/aggressive behavior (i.e., behavior characterized by subtle defiance and/or resistance slightly over the threshold of noncompliance), active resistant behavior (i.e., "actively [resisting] any form of problem resolution or arbitration" marked by decreased impulse control; Nicoletti et al., 2018, p. 52), and assault (i.e., behaviors which intend to harm). The most threatening form of physical abuse on this continuum is deadly assault, which occurs when an individual "focuses on killing a specific target, harming a group of individuals, or committing suicide" (Nicoletti et al., 2018, p. 52).

Intersectionality. Intersectionality emerged as an academic endeavor in the late 1980s in response to the judicial treatment of Black[3] women and women of color (Crenshaw, 1989, 1991). Although Black women activists have utilized "intersectionality-like thought" to fight overlapping forms of marginalization and discrimination since the early 19th century (Hancock, 2016, p. 24), it was Crenshaw (1989) who is frequently credited with coining the term (Iverson, 2017). Using the analogy of the four-way traffic intersection, Crenshaw (1989) explained:

> The point is that Black women can experience discrimination in any number of ways and that the contradiction arises from our assumptions that their claims of exclusion must be unidirectional. Consider an analogy to traffic in an intersection, coming and going in all four directions. Discrimination, like traffic through an intersection, may flow in one direction, and it may flow in another. If an accident happens in an intersection, it can be caused by cars traveling from any number of directions and, sometimes, from all of them.

[3] In her article, Crenshaw (1991) specifically noted that "Black" was capitalized because "Blacks, like Asians, Latinos, and other 'minorities,' constitute a specific cultural group and, as such, require denotation as a proper noun," whereas "white" is not capitalized "since whites do not constitute a specific cultural group" (p. 1244). "Women of color" are also not capitalized for the same reason.

> Similarly, if a Black woman is harmed because she is in the intersection, her injury could result from sex discrimination or race discrimination. (p. 149)

This analogy recognizes that Black women – and other women of color – experience discrimination distinctively from white women and Black men (Crenshaw, 1989). Crenshaw (1989, 1991) wrote that Black women are said to share discrimination experiences in ways similar to white women due to gender inequities and Black men due to racial inequities, but in reality, they experience discrimination in ways fully unique to them as Black women due to the combined effects of gender and racial inequities. As Crenshaw (1991) concluded, "the intersection of racism and sexism factors into Black women's lives in ways that cannot be captured wholly by looking at the race or gender dimensions of those experiences separately" (p. 1244).

Crenshaw (1989, 1991) used intersectionality to illustrate the ways Black women and women of color have been ignored by the judicial system, especially in response to forms of interpersonal violence. Part of this disregard is due to a mutually exclusive response by those fighting discrimination. As marginalized groups "organized against the almost routine violence that shapes their lives" (Crenshaw, 1991, p. 1241), the identity politics that emerged has failed to recognize intragroup differences. As Crenshaw (1991) noted, "Although racism and sexism readily intersect in the lives of real people, they seldom do in feminist and antiracist practices… Contemporary feminist and antiracist discourses have failed to consider intersectional identities such as women of color" (p. 1242). To better understand how women of color experience interpersonal violence, Crenshaw (1991) offered three perspectives of intersectionality to consider: structural, political, and representational.

Structural intersectionality describes "the ways in which the location of women of color at the intersection of race and gender makes [their] actual experience of domestic violence, rape, and remedial reform qualitatively different than that of white women" (Crenshaw, 1991, p. 1245). In other words, interpersonal violence for women of color has racial as well as gendered aspects (Linder, 2018), especially due to systemic policies and practices. For instance, Linder (2018) and Harris (2017) noted that women of color must navigate "racialized sexist oppression" (Linder, 2018, p. 17) based on gendered *and* racial stereotypes (e.g., Latina women as "hot and spicy" and Black women as promiscuous) that white women do not experience. These stereotypes can influence how women of color victims of sexual assault are viewed and victim-blamed more than white women victims, especially if the perpetrator is white (Donovan, 2007; George & Martinez, 2002). More broadly, structural intersectionality can also be used to understand how multiple social systems overlap to shape the experiences of those holding more than one marginalized identity in specific contexts, such as higher education (Museus & Griffin, 2011).

Political intersectionality, however, illustrates "how the multiple social groups to which an individual belongs pursue different political agendas, which can function to silence the voices of those who are at the intersection of those social groups" (Museus & Griffin, 2011, p. 7). Crenshaw (1991) used this perspective of intersectionality to explain how feminist and antiracist policies have further marginalized women of color when they experience interpersonal violence. For instance, the violence response systems in place on many college and university campuses rely on police officers as well as institutional judicial processes and policies (Iverson, 2017; Linder, 2018). As Linder (2018) noted, given the history of racism and sexism in policing as well as higher education, women of color who experience sexual violence on college and university campuses may not fully trust the reporting systems in place. Unlike their white women classmates, these women "must balance the tension between reporting experiences of sexual violence to police with their own safety and the safety of their male counterparts who may experience violence in the hands of police" (Linder, 2018, p. 18). Thus, political intersectionality interrogates the reality that while white women may argue for the use of policing to enact physical safety, these movements overlook the experiences of women who also identify as racial minorities.

Lastly, representational intersectionality refers to "the cultural construction of women of color" and how "the representation of women of color in popular culture can also elide the particular location of women of color, and thus become yet another source of intersectional disempowerment" (Crenshaw, 1991, p. 1245). For instance, racially themed parties hosted by white student organizations on college and university campuses can perpetuate stereotypes of underserved students and contribute to a negative campus climate for marginalized groups (Garcia et al., 2011). At these events, students are encouraged to "show up dressed representing racial stereotypes or to mock any racial or ethnic group" (Garcia et al., 2011, p. 6). In many cases, women students would ridicule the stereotypes of women of color, such as dressing as pregnant teenagers for a "South of the Border" themed-party or wearing blackface and padding their rear ends for a "Living the Dream" themed-party. In one particular instance at a Vietnam War themed party, women students attended dressed up like Vietnamese prostitutes "perpetuating stereotypes of Asian women as exotic and submissive" (Garcia et al., 2011, p. 8).

In the almost 30 years since Crenshaw first introduced intersectionality to the legal literature, scholars and theorists in other disciplines – such as history, sociology, literature, philosophy, education, and identity studies – have utilized it frame research, teaching, and social justice practice related to gender and race (Cho, Crenshaw & McCall, 2013; Collins, 2015). As Collins (2015) observed,

> Variations of intersectional scholarship can now be found across interdisciplinary fields as well as within more traditional disciplinary endeavors (Collins & Chepp 2013). Variations of intersectional practice can also be found within and outside the academy. Teachers, social workers, parents, policy advocates, university support staff, community organizers, clergy, lawyers, graduate students, nurses, and other practitioners find themselves upholding and challenging social inequalities. Practitioners both search for and propose ideas that will explain their experiences with the social problems around them. (pp. 2-3)

Scholars have also begun to use intersectionality to describe how overlapping structures of privilege and oppression operate for other social systems, including class, sexuality, ethnicity, nation, ability, and age, in addition to gender and race (Moradi & Grzanka, 2017). Moradi and Grzanka (2017) specifically challenged the notion that intersectionality "applies only to some people" and instead advocated for scholars to use "intersectionality to examine the breadth of experiences of social inequalities, including understudied axes of power, and experiences of privilege along with oppression" while still acknowledging intersectionality's foremothers (pp. 504-505). As such, intersectionality not only operates as theoretical lens to frame research by and about Black women and women of color, but as practical effort to improve the lives of people with multiple marginalized identities (Dill & Zambrana, 2009).

Intersectional scholars noted that to achieve these goals of social justice and social change, intersectional analysis must operate at both the individual level as well as the societal/structural level (Collins, 2015; Dill & Zambrana, 2009; Linder & Harris, 2017; Iverson, 2017). Linder and Harris (2017), for example, explained how using intersectionality can be used at both the microlevel and macrolevel to understand and prevent campus sexual violence: “This theory not only highlights the differing experiences of sexual violence for individuals with intersecting identities but also focuses on and destabilizes macrolevel systems of oppression that influence sexual violence on college campuses” (p. 242). These scholars noted that when higher education administrators use intersectional approaches at the microlevel, the experiences and voices of populations not usually discussed and represented as survivors of sexual assault (i.e., men, trans* students, women of color, students with disabilities, etc.) become centered and programming and policies for prevention and response change. Intersectionality challenges educators to focus on systems of overlapping identities instead of a single identity group, such as separating men and women for prevention programming (Linder & Harris, 2017). At the macrolevel, intersectionality illuminates how higher education in the United States continues to reproduce power structures that contribute to campus sexual violence (Linder & Harris, 2017). Institutional administrators will not end campus sexual violence unless they recognize and root out the many customs and traditions that propagate patriarchal, white supremacist, transphobic, heteronormative ideologies; it is the perpetuation of “these systems and institutions, not alcohol, hormones, or the way women dress” that influence “men’s feelings of privilege, power, and entitlement to violate and own others’ bodies” (Linder & Harris, 2017, p. 242).

Violence summary. The purpose of this section was to provide an overview of campus violence in order to frame the many contexts in which bystander intervention can occur. Despite the specific notions of campus violence put forth by the Clery Act – as physical crimes committed within a physical boundary – violence actually occurs along a continuum of verbal and physical abuse. Unfortunately, the culture and climate of many colleges and universities create an environment “at risk” for sexual misconduct and other forms of violence (McMahon, 2010, p. 3). The preponderance of “ordinary violence” (Cantalupo, 2009, p. 613) at institutions of higher education, coupled with the oftentimes inadequate responses from campus officials, means that “attending college is not a safe haven” for many students (Coker et al., 2011, p. 778).

Since violence has wide-spreading and long-lasting effects on a campus community, college and university administrators must continue to advocate for prevention measures as well as create a campus culture supportive of survivors and victims. They must also use microlevel and macrolevel intersectional approaches to understand and prevent all forms of campus violence. Iverson (2017), for example, analyzed the sexual violence policies of 22 institutions receiving U.S. Department of Justice’s (2015) Office of Violence Against Women (OVW) campus grants in 2012 and found that most institutions utilize “neutral” language that ignores the systems of power and privilege on campus. These “neutral” policies ultimately resulted in ineffective responses, leading Iverson (2017) to challenge “policymakers to move beyond institutionalized vocabularies and consider what individual and structural effects exist when sexual violence occurs at the intersections of ethnicity, language, and social class” (p. 227).

The activation of student bystanders to intervene in violent situations (or those with violence potential) is another approach campus administrations have used to prevent various forms of sexual violence. Although the training of bystanders on college campuses has focused on noticing and preventing this one type of campus violence, students also witness other forms of violence. To actually prevent campus violence, however, bystanders must understand their role and decide to intervene on behalf of the victim. In the following sections, I discuss how bystander decision-making as a cognitive process is currently theorized before offering a framework for conceptualizing how one develops a bystander identity.

Theoretical Perspectives of Bystander Decision-Making

Given the many disciplines that have explored bystander intervention and prosocial behaviors more generally (Dovidio et al., 2006), it comes as no surprise that scholars have also utilized a variety of approaches and frameworks to understand who bystanders are and how they respond. In this section, I offer three perspectives from differing scholarly traditions that each contribute to our overall understanding of bystander decision-making. The first framework, the Socioecological Developmental Model of Prosocial Action (Carlo & Randall, 2001) has foundations in Bronfenbrenner's (1979, 1993) ecology model and as well as social cognitive theories (Bandura, 1986). The Decision Model of Helping, which is most often used by social psychologists (see Dovidio et al., 2006; Latané & Darley, 1970), explores the meaning-making processes by which bystanders decide to intervene. Lastly, the third framework focuses specifically on bystander intervention as a question of morality and moral development: Is intervening the "right" thing to do? (see Batson, 1998; Hoffman, 2000).

Socioecological Developmental Model of Prosocial Action. Carlo and Randall (2001) used Bronfenbrenner's (1979, 1993) ecology model as the foundation for their Socioecological Developmental Model of Prosocial Action. This framework, which also draws from aspects from social cognitive theory (Bandura, 1986) and existing models of prosocial behavior (Eisenberg, 1986; Latané & Darley, 1970; Staub, 1979), "attempts to integrate the ecological, individual, and social and interpersonal influences of prosocial behavior" (Carlo & Randall, 2001, pp. 155-156). Specifically, Carlo and Randall focused on family and social contextual background variables, cognitive and emotive variables, and immediate situational characteristics to explain prosocial actions.

Carlo and Randall (2001) included family and social contextual background variables to acknowledge the important influences of family, peers, and culture on prosocial behaviors. For example, not only have researchers found that exposure to prosocial models in the media can enhance prosocial behaviors (Staub, 1979), but certain parenting styles are also associated with increased prosocial tendencies (Dekovic & Janssens, 1992). However, the empirical evidence at the time suggests that these variables are less likely to have a strong direct effect on prosocial tendencies. They instead indirectly influence the effects of other variables such as personal cognition or situational context on prosocial behaviors.

This model additionally accounts for an individual's cognitive and emotive traits and processes when examining their prosocial tendencies (Carlo & Randall, 2001). Included in this component are task-specific as well as global dispositions and skills, such as values, cognitive reasoning, self-efficacy, self-concept, attributions, perceptions, memory processes, perspective-taking, and moral reasoning. Perspective-taking and moral reasoning particularly influence prosocial behaviors since they most saliently facilitate an individual's understanding of and orientation to the needs of others (Underwood & Moore, 1982). High levels of empathy and sympathy are also associated with high levels of prosocial behavior (Eisenberg & Miller, 1987; Hoffman, 2000). Carlo and Randall (2001) noted that these factors directly influence one's prosocial behaviors in addition to shaping one's perceptions of the immediate situational context.

The third component of Carlo and Randall's (2001) model comprises of the immediate situational characteristics in which bystanders find themselves. Situational characteristics can range from the physical environment to the behaviors of other bystanders to the attributes of the victim (and possibly perpetrator) and include clarity of need, identity, physical attractiveness, and ease of escape (Batson, 1998). Despite the research suggesting that environmental context directly influences helping behaviors (see Latané & Darley, 1970), Carlo and Randall (2001) explain that “situational characteristics often have indirect and multiplicative, rather than direct and additive, effects. That is, immediate situational characteristics can facilitate or mitigate helping behaviors usually through their impact on cognitive and emotive traits and processes” (p. 158).

Bronfenbrenner’s influence on this model becomes most apparent, however, through the inclusion of the feedback process depicting how past prosocial behaviors impact family and social contextual background variables as well as an individual’s cognitive and emotive traits and processes, thus influencing future prosocial behaviors (Carlo & Randall, 2001). For example, some research suggests that prosocial actions can lead to an individual developing an identity as a moral or benevolent person (Eisenberg & Fabes, 1998). A few “cross-cultural studies” (see Eisenberg & Fabes, 1998) have also demonstrated that “collectivist societies encourage and promote frequent acts of cooperative and prosocial behaviors” through group norms and expectations that reward prosocial behaviors (Carlo & Randall, 2001, p. 159). However, Carlo and Randall recognized the need for more research in this area to better understand how this feedback process transpires and the specific variables required, particularly when it involves diverse populations.

This social-ecological framework provides a big-picture perspective of the factors influencing bystander decision-making processes. Although it does not explicitly inform the specific steps and factors contributing to bystander decision-making in the moment, this model equips us with a framework for interrogating the broader environmental factors, including past bystander behaviors, relationships with others, organizational policies, community standards, and cultural norms, leading to bystander decision making. Where this perspective falls short, however, is explicit recognition of how these environmental contexts differ for bystanders from diverse backgrounds or for bystanders engaging in interactions across differences. Given our systemic national issues around race, religion, gender, sexual orientation, and nationality, further research is needed to understand how our differences influence bystander intervention at all levels.

The decision model of bystander helping. Training community members to act as prosocial bystanders has become a common solution to sexual violence on college and university campuses because:

> The social nature of sexual violence presents opportunities for bystanders to act and intervene to prevent sexual violence and to be supportive of victims. Despite the social nature of sexual violence, that it is often consigned to a private realm that makes it difficult for friends and loved ones to respond. However, victims, if they tell anyone, will tell a friend; it is important that friends be taught to recognize the signs and to intervene. (Amar, Sutherland, & Kesler, 2012, p. 851)

Although bystander intervention is traditionally referenced within the context of emergencies or extreme violence, there exists a whole range of possible violent scenarios in which bystanders can intervene (Nelson, Dunn, & Paradies, 2011). In fact, bystanders are more likely to witness "ordinary violence" than crisis situations (Cantalupo, 2009, p. 613; Zoccola et al., 2011). If and when a bystander decides to intervene in any situation is complex (Burn, 2009; Dovidio et al., 2006). Latané and Darley (1970) first attempted to answer this question with their five-step process. Although this framework was initially developed to understand bystander behavior in emergencies that require immediate assistance, scholars have found this model also applies to intervening in ordinary or common situations (see Dovidio et al., 2006). In this model, bystanders must first notice the event (step 1) and then interpret it as a situation requiring intervention (step 2). They must next decide to take responsibility for acting (step 3) and decide how to act (step 4), before finally choosing to act (step 5). At any stage in this process, they may come across barriers, or inhibiting factors, which prevent them from intervening on behalf of the victim (Burn, 2009; Dovidio et al., 2006; Latané & Darley, 1970).

The most common barriers at the first step – noticing the event – include self-focus or distractions (Burn, 2009). For example, patrons at a bar may fail to notice a potential sexual assault because they are either focused on their social activities or are intoxicated themselves, therefore compromising their cognitive processes. Bystanders may also miss instances of harassment in their residential environment because they are in a rush to get to work or are focused on their lives. Given the number of activities occurring on college campuses and self-focus of college students, it is not surprising that many students do not notice instances of harm, or potential harm, which prevents them from moving to the other steps of acting as a prosocial bystander.

Once bystanders notice a situation, they must also decide it requires intervention (Step 2; Latané & Darley, 1970). Burn (2009) identified ambiguity and ignorance as inhibiting factors at this step; the less ambiguous the situation, or the more knowledgeable the bystanders, the more likely someone is to intervene (Fischer, Greitmeyer, Pollozek, & Frey, 2005; Harari, Harari, & White, 2001). These barriers are most common for low-risk types of violence. For example, when a bystander overhears a derogatory comment made about members of a particular community, they may recognize it as wrong, but not as something worthy of intervention due to the ambiguity of risk. Ambiguous situations also inhibit bystander behavior due to “pluralistic ignorance” (Clark & Word, 1974; Darley & Latané, 1968), which is when “ignorant, inactive bystanders look to other ignorant, inactive bystanders and consequently all fail to identify the situation as intervention appropriate" (Burn, 2009, p. 781). For example, bystanders may not know the risk markers of sexual assault, or what constitutes consent, when they allow their intoxicated friend to leave a party with a stranger (Burn, 2009).

After bystanders recognize a situation as one which requires intervention, they are more likely to act if they feel it is their responsibility to help (Step 3; Latané & Darley, 1970). Failure to take responsibility is most likely influenced by three factors: the presence of others, the relationship of the bystander to those involved in the situation (either the victim, the perpetrator, or both as well as other bystanders), and beliefs about the potential victim's worthiness for assistance (Burn, 2009). The presence of others decreases intervention behaviors because it diffuses the responsibility of taking action among those available to help; bystanders assume action is not their responsibility because someone else will handle it (Brody & Vangelisti, 2016; Chekroun & Brauer, 2002; Darley & Latané, 1968). Diffusion of responsibility is what most scholars believe caused of inaction by Kitty Genovese's neighbors in 1964 (Dovidio et al., 2006).

Bystander anonymity (i.e., no one in the situation knowing the bystander) also negatively influences their responses in emergencies and situations of interpersonal violence (Brody & Vangelisti, 2016; Schwartz & Gottlieb, 1980; Solomon, Solomon, & Maiorca, 1982). Early research suggests that bystanders who believe they are anonymous to other bystanders will not intervene (Schwartz & Gottlieb, 1980; Solomon et al., 1982). Schwartz and Gottlieb (1980) hypothesized that social norms influence this effect; in their study, "not a single anonymous bystander in the presence of another reported thinking that others had any expectations of them, whereas 23% of the known bystanders explicitly indicated 'some sense that as another knew that

I had witnessed the attack, I was expected to act'" (p. 423). More recently, Brody and Vanelisti (2016) found that perception of online visibility is positively related to passive observing by undergraduate students who witnessed a case of cyberbullying on Facebook. In other words, undergraduate students were more likely to do nothing in response to hurtful online comments if they thought no one could see them. Brody and Vanelisti (2016) also suspected the influence of perceived anonymity on reducing the regard for social standards as the reason for this outcome.

The relationship of the bystander to those involved also influences the level of responsibility they feel toward intervention, but how is still up for debate. Some scholars believe that responsibility to intervene increases if the bystander has a relationship with the victim, or potential victim (Brody & Vangelisti, 2016; Gottlieb & Carver, 1980; Howard & Crano, 1974; Levine, Cassidy, Brazier, & Reicher, 2002), while others have evidence that knowing the victim decreases likelihood to intervene (Dahl, Lo, Youngerman, & Mayhew, 2017). Banyard (2008), on the other hand, found no relationship between knowing the victim and bystander behavior. For example, the bystanders who allow their intoxicated friend to leave a party with a stranger may not think it is their responsibility to intervene because they do not want their friend to get mad at them (Dahl et al., 2017). Knowing other people involved, such as the perpetrator, also influences bystander behaviors. In other words, multiple bystanders who witness the same event at the party described above may know the perpetrator instead of the victim and believe this person could be doing the right thing, even if he or she is not.

Victim worthiness also contributes to bystanders' sense of responsibility (Loewenstien & Small, 2007). Alcohol use, previous behavior, and provocative attire all decrease bystanders' perception of victim worthiness in instances of sexual assault and cyberbullying (Cassidy & Hurrel, 1995; Norris & Cubbins, 1992; Schult & Schneider, 1991; Shultz, Heilman, & Hart, 2014; Whatley, 2005; Workman & Freeburn, 1999). Additionally, in cases of sexual assault where both the offender and the victim are equally intoxicated, bystanders are more likely to hold the victim more responsible for the act than the perpetrator (Abbey et al., 1996; Castello, Coomer, Stillwell, & Cate, 2006; Fogle, 2000; Sampson, 2003). Acceptance of rape myths, such as “a drunk person was asking for it” or “drunk people can’t control their sexual impulses,” not only contributes to perpetration of sexual violence (Bohner, Jarvis, Eyssel, & Siebler, 2005; O’Donohue, Yeater, & Fanetti, 2003), but also devalues a victim’s, or potential victim’s, worthiness of help (McMahon, 2010).

The fourth step bystanders must take to intervene is to decide how to help (Latané & Darley, 1970). Many bystanders who get to step four do not intervene because they do not feel as though they have the skills (Burn, 2009; Cramer, McMaster, Bartell, & Dragna, 1988; Shotland & Heinold, 1985) and lack the confidence in their ability to intervene effectively (Burn, 2009; Goldman & Harlow, 1993; Latané & Darley, 1970). Banyard (2008) found that self-perception of bystander efficacy to intervene in events of interpersonal violence were positively associated with bystander behavior, indicating that bystanders with higher self-efficacy are more likely to intervene on behalf of others. Additionally, students are also more likely to act as a prosocial bystander if they have seen someone else model the behavior first (Bryan & Test, 1967; Rushton & Campbell, 1977).

Once bystanders complete step four, “all that remains is the actual act” (Step 5; Latané & Darley, 1970, p. 35). Although most bystanders will have little difficulty implementing their decision from step four, they may still be impeded by fears of embarrassment and awkwardness (i.e., “audience inhibition”) even if they do not know anyone in the situation or other bystanders (Burn, 2009; Latané & Darley, 1970; McMahon & Dick, 2011). This anxiety toward the thought of possible negative evaluations or actions by others prevents action because bystanders fear they will make a mistake (Burn, 2009). For example, bystanders overhearing harassing comments directed at another person may not intervene because they are afraid of making the situation worse, even if they have already decided how to help. The number of bystanders can also elevate a bystander's apprehension, demonstrating that "audience inhibition may reduce bystander intervention at large parties or in bars unless there are salient social norms consistent with intervention" (Burn, 2009, p. 782). Due to the prevalence of others, alternative campus contexts where audience inhibition may become a factor include classrooms, residence hall programs, sporting events, and student organization meetings.

Although this model has provided many scholars with an analytical framework for understanding bystander behaviors (see Fischer et al., 2011), it does not present a complete picture of prosocial decision-making (Dovidio et al., 2006). For example, what about those situations in which the need is apparent and the focus of responsibility is clear, but intervening poses a great danger to the bystander? Do bystanders still act on behalf of the other, or do they practice self-preservation? How and why do bystanders come to this conclusion? An additional limitation to this model is consideration for diverse populations. How does race or ethnicity influence helping decision-making? What about religious perspective and worldview? Or gender identity? Or sexual orientation? And how do these factors influence decision-making within various contexts given the power and privilege granted to those with dominant identities?

As a response to the limitations of Latané and Darley's (1970) framework, Jane Piliavin and her colleagues (1981) integrated a cost-reward perspective with the decision model to explain the influence of costs and rewards on bystander helping behaviors. This approach posits that in a potential helping situation, "a person analyzes the circumstances, weights the probable costs and rewards of alternative courses of action, and then arrives at a decision that will result in the best personal outcome" (Dovidio et al., 2006, p. 85). Costs to the bystander for assisting include effort and time (i.e., the interruption or postponement of something important), potential personal harm, psychological aversion (i.e., the situation involves something unpleasant), financial expenditure, or social disapproval (i.e., the situation challenges a social norm). Rewards for helping, however, may bring monetary compensation and social benefits such as fame, gratitude, and reciprocity. Yet, what if the bystander does not help? They could then experience guilt, blame, or other unpleasant feelings associated with witnessing suffering, not to mention, the victim could experience serious harm (Piliavin et al., 1981).

The influence of social norms on this cost-reward analysis is particularly salient for college students who find themselves in bystander situations. For example, in their study of college students' past bystander behaviors related to gender prejudice, Brinkman, Dean, Simpson, McGinley, and Rosén (2015) found that concerns about social norms were more likely to predict unutilized prosocial responses for women than men, although both groups appeared to use prosocial responses at similar rates. This gendered difference indicates that women more often consider intervention but hold back because they are concerned they will become the target of prejudice themselves or will be treated badly by other bystanders for intervening, a perceived as negative consequence for breaking social norms. Furthermore, Carlson (2008) found that social norms around "masculinity may be another factor in the complex behavior of bystanders to violent situations" (p. 13). Specifically, the men in this study felt strongly that they must not appear weak to others, especially other men. This perspective often influenced other behaviors such as drinking and fighting as well as decision-making around intervening in harmful situations. These men considered intervention in public settings with both men and women present masculine, but intervening in a private setting with “just the guys” weak and unmasculine.

As bystanders consider these costs and rewards, the nature of the relationship between them and the victim, as well as perceptions of deservingness and clarity of need, also influences their decision (Dovidio et al., 2006). Several researchers have found that interpersonal attraction, whether based on physical appearance, friendly behavior, or other personal qualities, can increase helping behaviors since there is the possibility of increased potential rewards (Dovidio & Gaertner, 1983; Dovidio et al., 2006; Harrell, 1978; Kelley & Byrne, 1976; Kleinke, 1977). People also tend to intervene on behalf of others who appear similar to themselves (Dovidio, 1984) or with whom they share group membership (see Dovidio et al., 2006). In both of these contexts, the perceived rewards outweigh the potential costs. Guilt can be higher for not helping a member of your group; so, too, can appreciation for assisting (Dovidio et al., 2006). Banyard, Weber, Grych, and Hamby (2016), for example, demonstrated that individuals with higher perceptions of community and microsystemic support (e.g., support of community youth, informal community support, and social support) tended to recognize bystanders as helpful in potentially harmful situations, which could contribute to prosocial behavior in future situations. Additionally, Abbott and Cameron (2014), found that among British adolescents, intergroup (i.e., interethnic) contact was positively associated with prosocial bystander intentions to intervene in an intergroup harassment scenario. These authors also found that empathy, cultural openness, and in-group bias also influenced the student’s prosocial intentions.

When it comes to racial differences and helping in the United States, decision-making processes become more complicated. Experimental research from the 1970 and 1980s suggests that white participants will help those they believe belong to other racial groups less often or less quickly than other white persons in need (Benson et al., 1976; Gaertner, Dovidio, & Johnson, 1982), yet more recent research suggests mixed results. Although Saucier, Miller and Doucet (2005) found no differences associated with race and helping, results from Kunstman and Plant's (2008) research on the influence of emergency severity and racial bias in helping behavior indicate that severe emergencies involving Black victims elicited higher levels of aversion and decreased speed and quality of help from white participants, relative to White victims.

Although these changes in behavior by whites may be due to evolving cultural perceptions of race, Dovidio and Gaertner (2004) hypothesized it may occur because of aversive racism. Aversive racism represents a subtler form of racial bias in which "Whites who may truly believe they are not prejudiced still harbor unconscious negative feelings toward Blacks" (Dovidio et al., 2006, p. 97). These "unconscious negative feelings" are one reason why white participants were more likely to help Black persons in need only if it seemed no one else could help, if it would not take too much time, or if the help seemed not too difficult or requiring significant effort, all costs which could be justified as not necessarily associated with race (Dovidio et al., 2006; Saucier et al., 2005). For example, white participants in Kunstman and Plant's (2008) study interpreted an emergency with a Black victim as less severe, and thus, themselves less responsible to help, than an emergency with a white victim. Averse racism additionally offers an explanation as to why "emergency racial bias is unique to white individuals' responses to Black victims and not evinced by Black helpers" (Kunstman & Plant, 2008, p. 1499).

More recent research by Katz, Merrilees, Hoxmeier, and Motisi (2017) and Katz, Merrilees, LaRose, and Edgington (2018) suggest that aversive racism influences white women collegians' decisions not to help their Black women peers when they are at risk for incapacitated sexual assault. In both of these studies, participants reported less intent to intervene and less personal responsibility to intervene when the victim was thought to be a Black woman than when the victim's race was ambiguous. Additionally, the first study (Katz et al., 2017) noted that the white women in the study reported greater perceived victim pleasure when the victim was thought to be Black. The second study (Katz et al., 2018) found that sexist attitudes were related to increased blame and reduced willingness to intervene on behalf of Black women and that these adverse effects were reduced by concerns about racial injustice.

As bystanders balance the possible costs with the potential rewards in their decision-making, they also decide how to assist (Piliavin et al., 1981). If the victim's costs for not receiving help are high, but the bystander's costs for assisting are low, then bystanders will most likely use direct intervention to help the person in need. If the bystander perceives the costs for helping as high (e.g., potential injury, effort, or embarrassment), they will resort to either indirect intervention (i.e., seeking someone else with more authority to help) or will redefine their perception of the situation in order to feel better about ignoring it (e.g., diffusion of responsibility, disparagement of the victim, etc.). Additionally, the victim's cost for not receiving help can also be low. In this case, Piliavin et al. (1981) posited bystanders will either ignore or deny the need for help and not do anything if their perceived costs are high, or they will rely on the norms associated with the situation if their perceived costs are low.

Latané and Darley's (1970) model of bystander intervention behavior was first published over 45 years ago and it continues to represent, in part, the decision-making process modern bystanders use (see Burn, 2009; Dovidio et al., 2006; Fischer et al., 2011). Bystanders still must notice the event, interpret it as requiring assistance, assume personal responsibility, choose a way to help, and act to provide help. However, the original framework fails to reflect the many costs and rewards bystanders also consider when deciding to help, even after assuming personal responsibility. Piliavin et al.'s (1981) addition of the cost-reward analysis to the framework offered a "powerful refinement" to Latané and Darley's model (Dovidio et al., 2006, p. 103). This adjustment allows for further examination of not only the psychological processes used by bystanders to evaluate a situation and subsequent actions, but also considers how context and social norms play an important role in bystander behaviors, echoing the socio-ecological perspectives presented above.

Although this model provides a thorough explanation for when bystanders will intervene, one limitation is that it does not consider the influence of moral reasoning and development. The cost-reward analysis of helping assumes people are motivated to maximize rewards and minimize costs (Dovidio et al., 2006), yet humans do not always reason in an economically rational way (March & Olsen, 1980). The following section provides an overview of the third framework for understanding bystander intervention: morality and moral reasoning.

Theories of bystander moral development. Theories of moral reasoning and development provide an additional framework for understanding who bystanders are and what bystander intervention means (Batson, 1998). As Carlo & Randall (2001) noted:

> Moral reasoning refers to thinking in dilemma situations where issues of justice, fairness, or caring are prevalent. Often, an individual's style of moral reasoning reflects an orientation to the needs of others or their own. Furthermore, individuals' preference for some types of moral reasoning is linked to values or emotions (e.g., sympathy) that facilitate responding to others' needs. (p. 157)

Scholars conceptualize the development of moral reasoning in several ways, such as progression on a "standard scale of moral rightness" in which morality is defined by an external authority

(Mayhew et al., 2016, p. 332), the process by which individuals move from “simple and finite” perspectives of right and wrong to more complex ways of reasoning (Dorough, 2011, p. 59), or a synthesis of empathic affect and the advancement of a cognitive sense of others distinct from the self (Hoffman, 2000). Regardless of perspective, as bystanders advance in their moral reasoning, their prosocial behaviors increase (Underwood & Moore, 1982). For instance, one would expect that bystanders with “more complex ways of distinguishing right from wrong” (Dorough, 2011, p. 59) would be more willing to take responsibility and know which actions are best suited to the situation, and thus, be more likely to intervene (Carlo & Randall, 2001).

Development of moral reasoning has long been considered one of the central outcomes of participation in higher education (Mayhew et al., 2016). Within the context of higher education research and theories of college student development, Lawrence Kohlberg, Carol Gilligan, and James Rest prevail as the three primary scholars of moral reasoning development. Although these theories provide an additional framework for understanding how and why bystanders decide to take action – or inaction – on behalf of others, they were not explicitly developed with empathy or harm in mind. Martin Hoffman, however, has focused on both aspects in developing his theory of prosocial moral behavior and development. This theory, which “highlights empathy’s contribution to moral emotion, motivation, and behavior” (Hoffman, 2000, p. 3), also considers cognition as well as principles of caring and justice, which are aspects of other theories of moral reasoning. By considering empathy’s role in moral reasoning, he examined how individuals resolve caring-justice conflicts in moral dilemmas.

Psychologists define empathy in two ways: (1) the cognitive awareness of another person’s internal states; or (2) the vicarious affective response to another person. Hoffman’s (2000) theory focused on empathy and moral reasoning in five types of dilemmas in the prosocial moral domain: innocent bystanders, transgressors, virtual transgressors, multiple moral claimants, and caring versus justice. The moral issues facing these dilemmas include refraining from harming others, deciding who to help when others could potentially be neglected, and determining whether to choose justice over caring (or caring over justice). For instance, innocent bystanders, which Hoffman described as the prototypic moral encounter, must make the decision to help when witnessing someone else in physical, emotional, or financial pain or distress. Whether individuals are motivated to help, and if they do, to what extent is this motivation based on genuine concern for the victim is the moral issue for bystanders.

Hoffman's (2000) theoretical framework of moral reasoning also centers the development of empathic distress. This approach conceptualizes empathic distress as a process of empathic synthesis and cognitive sense of self which occurs in five stages: (1) reactive newborn cry, (2) egocentric empathic distress, (3) quasi-egocentric empathic distress, (4) veridical empathic distress, and (5) empathy for another's experience beyond the immediate situation. As with Kohlberg's theory, each stage in Hoffman's scheme builds upon the gains of the previous stages. Individuals in stage two (egocentric empathic distress) respond to another's distress as though they were suffering. In stage three (quasi-egocentric empathic distress), however, individuals realize that the distress of others is not the same as their own, yet they still respond by doing for the other what would comfort themselves. Once they reach stage four (veridical empathic distress), individuals realize that others are fully independent of themselves and are closer to feeling what the other is feeling, not what they think the other is feeling. Finally, individuals at stage five (empathy for another's experience beyond the immediate situation) can empathize with an entire group because they realize that others experience happiness and sadness. At this most advanced stage of empathic synthesis, "observers may act out in their minds the emotions and experiences suggested by [verbal and nonverbal expressions from the victim and situational cues] and introspect on all of it" (Hoffman, 2000, p. 7).

In addition to the stages of empathic synthesis, Hoffman (2000) also described how empathic distress is shaped into four empathy-based moral affects formed by the attributes of the distressing event. These affects – sympathetic distress, empathy-based anger, empathy-based feeling of injustice, and guilt over inaction – function as motives for individuals facing moral dilemmas. Sympathetic distress occurs when the cause of another's distress is beyond their control. Empathic anger happens when someone else is the cause of another's distress and can consist of either empathy with the victim's anger or simultaneous empathic sadness and anger at the perpetrator. When a discrepancy exists between the victim's character and their fate (i.e., a good person experiences something bad), the observer feels an empathic feeling of injustice. If, however, the observer feels as though the victim is deserving of this fate, the observer will blame the victim for their own suffering. Finally, individuals feel guilt over inaction if they do not help (no matter how legitimate the reasons) or if their efforts fail to help, which continues to cause distress to the victim. An important point about individuals' responses to empathic distress is that the victim is not required to be physically present; to feel this way, or any one of the moral affects, an individual need only to imagine the victims when learning about situations of hardship.

For bystanders, empathic distress is a prosocial motive to assist others in distress, but it does not always lead to helping. Hoffman (2000) posited several reasons for inaction, including pluralistic ignorance and diffusion of responsibility (as hypothesized by Latané & Darley, 1970). He also recognized, much like Pilivian et al. (1981), that bystanders evaluate the costs of helping, including egoistic feelings of fear, energy expenditure, financial cost, loss of time, and the unpleasantness of experiencing empathic distress. In short, Hoffman's bystander model "involves conflict between the motive to help and egoistic motives that can be powerful" (p. 35) as individuals make the empathic moral decision to intervene.

As hypothesized above, a clear connection exists between the Pilivian et al.'s (1981) refined model for bystander decision-making (Latané & Darley, 1970) and theories of moral reasoning development and behaviors (see Batson, 1998; Carlo & Randall, 2001). Hoffman (2000) explicitly considered how empathy and empathy-based moral affects influence these behaviors. A bystander may exhibit empathy for another's experience beyond the immediate situation (stage five) and feel empathic anger for a victim's situation but could decide not to intervene because of potential opportunities missed or conflicting moral stances of helping and justice. For instance, a student who strongly believes abortion is wrong for moral and religious reasons could witness a pro-choice protestor in a harmful stance-related situation. This student, who feels empathic for the protestor and anger at the perpetrator in the situation, may not intervene because they do not want others to see them as empathic to the protestor's cause; the student may also be unsure if helping the protestor conflicts with their views of justice related to abortion.

Although most student bystander intervention scholars have yet to frame their work using theories of moral reasoning development, these perspectives provide an excellent approach to understanding how bystanders make meaning of harmful situations beyond that of costs and rewards. Hoffman's (1970, 2000) theory provides an additional perspective which includes aspects of empathic distress and empathy-based moral affects also experienced by bystanders to harmful events. Despite the presumed functionality of these theories, little is known about their applicability across diverse groups and experiences. Exacerbating this matter is the lack of scholarship exploring how differences in identity influence moral development, particularly for college students (Mayhew et al., 2016). In the last two decades, only three studies have attempted to measure these types of conditional effects on moral reasoning, with only one significant interaction effect being reported: the effect of diversity-based coursework on moral reasoning development varied for students from dissimilar income levels (Bowman, 2009). Specifically, Bowman discovered that affluent students who enrolled in two diversity courses experienced greater gains in their moral reasoning development than lower-income students, yet once students enrolled in three or more courses, the low-income students benefitted more than their affluent counterparts.

Bystander decision-making summary. Scholars have used several different approaches to guide the understanding of bystander intervention. Socio-ecological models provide scholars with a way to make sense of the many contextual factors influencing bystanders' perceptions of a situation and appropriate responses. Yet, the socio-ecological models are limited in their ability to describe the cognitive processes by which bystanders interpret and act in distressing situations. The decision model of helping developed by Latané and Darley (1970), and refined by Piliavin et al. (1981), catalogues the cognitive steps bystanders must complete, and the factors they consider, to successfully intervene. Despite the continued success of this model (see Fischer et al., 2011), it over-emphasizes the assumption of rationality and insufficiently considers the role of moral reasoning in human decision-making. Theories of moral reasoning, particularly Hoffman's (2000) theory of empathic morality, equip scholars with yet another framework for understanding how bystander emotions influence their actions. Each of the frameworks presented here adds to the understanding of who bystanders are and what bystander intervention means. In the final part of this chapter, I apply these models to inform my theoretical understanding of bystanders and bystander intervention in collegiate contexts.

Theoretical Perspectives of Bystander Identity Development

How college student bystanders perceive themselves and their social identities also influences their intervention behaviors in a way that has yet to be fully addressed by the bystander literature. However, this perspective needs to be included since "an understanding of identity is necessary if one is to understand college student and their experiences in higher education contexts" (Jones & Abes, 2013, p. 19). As Lewin (1936) noted in his equation, behavior is a function of the person and their environment; as such we should study the person as well as their environment in order to understand their behavior. It is for this reason that I would like to introduce the Reconceptualized Model of Multiple Dimensions of Identity (RMMDI; see Jones & Abes, 2013) as a framework for exploring how bystanders' social identities inform their perceptions of a situation. This model also offers a way to understand how bystanders can develop a bystander identity.

So far in this discussion, I have introduced the cognitive and moral reasoning processes by which college student bystanders come to their intervention decisions and behaviors. Despite their benefits, these frameworks have a few limitations. First, they fail to articulate the influence of helping self-efficacy and skills on decision-making and behavior. Latané and Darley (1970) addressed the impact of these traits somewhat in their discussion of barriers to helping as an individual arrives at step 5 of their model, but more detail is needed to fully understand the interaction of contexts and confidence in skills on bystander behavior. Second, and probably most important, they do not address the fact that bystanders of diverse backgrounds could use different decision-making and meaning-making processes than was initially developed. Context is essential here, as college students of marginalized backgrounds experience their campus cultures much differently than students holding majority identities (Mayhew et al., 2016). These differences matter as bystanders witness negative behavior and decide what actions to take.

Although social identities such as race and gender have been briefly explored as part of the cost-reward perspective of bystander decision-making due to their influence on social norms and perceived relationship to the victim, psychosocial development generally, and identity development specifically, have yet to be investigated as important aspects of college student bystander intervention disposition. The purpose of this section is to describe the RMMDI and my rationale for using it to understand how identity development and meaning-making influence college student bystander intervention.

Social identities. Before describing the RMMDI, I would like to first define concepts related to the model, namely the social construction of identities, privilege and oppression, and identity salience (Jones & Abes, 2013). According to Jones and Abes (2013), the first use of the term *social identity* is attributed to Henri Tajfel (1982), who conceptualized social identity as "that part of the individuals' self-concept [personal identity] which derives from their knowledge of their membership in a social group (or groups) together with the value and emotional significant attached to that membership" (p. 2; as cited in Jones & Abes, 2013, p. 36). This perspective suggests a relationship between one's perception of self and the salience of membership in various social groups. As Jones and Abes (2013) concluded, there is an "inextricable link between personal and social identities, between the individual, the social world, and the meaning the individual makes of his or her experiences" (p. 37). However, it should be noted that social identities are socially constructed; how one perceives their sense of self in relation to their social identities is constructed through interactions with others in the "broader social context in which dominant values dictate norms and expectations" (Torres, Jones, & Renn, 2009, p. 577).

This social construction of identities cannot be fully understood without also recognizing the "mutually reinforcing" roles of privilege and oppression (Jones & Abes, 2013, pp. 38-39). Privilege refers to the ways in which members of certain social identity groups receive systematic empowerment and entitlements not available to all groups of people (Johnson, 2006). One important aspect of privilege is the relative ease with which privileged groups can be unaware of how privilege affects them. White privilege, for instance, is why whites do not need to consider other racial groups, whereas "African Americans, for example, have to pay close attention to whites and white culture and get to know them well enough to avoid displeasing them, since whites control jobs, schools, the police, and most other resources and sources of power" (Johnson, 2006, p. 22). Oppression, on the other hand, results from the systematic holding back of people because of their membership in non-privileged identity groups; "Just as privilege tends to open doors of opportunity, oppression tends to slam them shut" (Johnson, 2006, p. 38). It is important to note, however, that although individuals can vary in how they perceive oppression, they must belong to an oppressed group in order to experience oppression at all. In other words, just like privilege is systematic based on the relationship between social categories, so is oppression (Johnson, 2006). These concepts are important to understanding identity development because "social identities are influenced by social constructions that emerge from structures of privilege and oppression. The complex ways in which privileged and oppressed identities intersect have an impact on individual perceptions of self and the identity construction process" (Jones & Abes, 2013, p. 40).

Finally, one's various social identities can have differing levels of prominence or importance based on the context, with individuals reconsidering their level of identification with certain social groups as their context changes (Ethier & Deaux, 1994, as cited in Jones & Abes, 2013). Ethier and Deaux (1994) suggested several ways in which context and salience influence one's social identity. First, some individuals have strong identification with a group, regardless of the context. For others, however, identities can become more or less salient when there exists an incongruity between their self-perceived social identity and their current context. This distinction can also occur when past contexts and experiences do not align with the current context, causing a change in identity salience. Furthermore, it should be noted that identity salience does not necessarily occur for one identity at a time. In fact, intersections of identities can simultaneously be salient or not given a particular context (Jones & Abes, 2013).

Overview of the RMMDI. Based on the original Model of Multiple Dimensions of Identity (MMDI) developed by Jones and McEwen (2000), the RMMDI reconceptualizes how meaning-making (Keegan, 1994) and self-authorship (Baxter Magolda, 2001, 2009) influences identity development (Jones & Abes, 2013). At the heart of these theories is the idea that identity occurs at the intersection of context, personal characteristics, and belonging to multiple social identity groups. The MMDI and RMMDI both contain several fundamental elements which create a dynamic understanding of identity when combined. These components include the core, multiple social identities, the relationship of social identities to the core and identity salience, and contextual influences (Jones & Abes, 2013). As with most theories of human development, this model is fluid and allows for the components to change as contexts and identity salience shift.

At the center of the multiple layers of identity is the core (Jones & Abes, 2013). The core represents an individual's "internal sense of self," which is why it appears at the center of the model (Jones & Abes, 2013, p. 82). For Jones and Abes (2013), a person's core is that part of their identity that is "impenetrable and protected from outside influence" (p. 82); it cannot be labeled by others as social identities are. In other words, the core is one's inside self, the aspect of one's perception of self that has the most agency and is the most stable.

Surrounding the core are the many intersecting social identities one holds (Jones & Abes, 2013). Jones and Abes (2013) emphasized a clear distinction between one's personal identity – the core – and the multiple social identities he or she can hold, which are socially constructed. By focusing on social identities in this way, the model requires scholars to acknowledge the many intersecting social systems and structures that contribute to privilege and oppression and the influence of these forces on identity development. As Jones and Abes (2013) remarked, "the process of coming to know oneself and thinking about the question 'Who am I?' is complicated by the socially constructed identities of race, gender, cultural, and sexual orientation, and their intersections" (p. 84). Additionally, by conceptualizing social identities as intersecting circles around one's core, the model allows for a more fluid and dynamic understanding of how social identities are negotiated based on the context.

The illustration of social identities as circles around the core also captures the relationship social identities have with the core through identity salience (Jones & Abes, 2013). When someone's social identities are more salient, they have closer proximity to the core; those identities with less salience are farther from the core. As these identities interact and salience shifts around the core, they are also influenced by systematic privilege and oppression. For most people, "systems of privilege and inequality [are] least understood by those who [are] most privileged by these systems. The more privileged an identity (for example, race), the less salient it [is]" (Jones & Abes, 2013, p. 85). On the flip side, one's experiences of difference and feelings of otherness can increase identity salience depending on the social identity's level of visibility (e.g., race or ethnicity versus social class or sexual orientation; Jones & Abes, 2013). This "prism of privilege and difference" ultimately mediates the connection individuals have with certain social identities and their relative salience (Jones & Abes, 2013, p. 86).

Finally, the MMDI and RMMDI also consider how one's core identity and multiple social identities are situated within a larger context (Jones & Abes, 2013). As Jones and Abes (2013) explained,

> The intent of nesting social identities within context is both to suggest that self-perceived personal and social identities may not by fully understood without considering larger external forces as well as to draw attention to particular contextual influences that made a difference to the participants in the original study on which the MMDI is based. (p. 88)

Context influences all the other aspects of the model, including one's identity salience, the ways multiple social identities interact and intersect, and experiences with privilege and oppression; one's identity development is "deeply embedded in and created out of contexts" (Jones & Abes, 2013, p. 88). Additionally, just as privilege is not always apparent to those with privileged social identities, the influence of context is not always perceptible to individuals who need not think about it. This awareness may be related to the intersection of one's social identities and contexts, one's "cognitive capacity for recognizing dimension of context," or the sheer fact that some contexts are simply indistinguishable (Jones & Abes, 2013, p. 90).

Although the MMDI has many strengths, one of its major limitations is its emphasis on only one domain of development: identity. Yet identity development does not occur exclusive from other forms of development, such as cognitive and interpersonal development. As such, Abes, Jones, and McEwen (2007) reconceptualized the MMDI to incorporate meaning-making processes. In this new model – the RMMDI – meaning-making capacities as defined by Kegan (1982, 1994) and Baxter Magolda (2001, 2009) now act as a filter through which students interpret the influence of context on their personal and social identities (Jones & Abes, 2013). In other words, the impact of contextual factors on self-perception and social identity salience depends on the complexity of one's meaning-making capacity.

Kegan's (1982, 1994) theory describing the evolution of consciousness focuses on how people "construct meaning" with respect to their life experiences (Kegan, 1994, p. 190). Meaning making consists of cognitive, intrapersonal, and interpersonal components. Central to this theory is the continual shifting between periods of stability and instability that marks one's evolution of meaning and leads to "ongoing reconstruction of the relationship of persons with their environments" (Evans et al., 2010, p. 177). In essence, development in meaning making results from one's effort to resolve the cognitive tension between a longing for distinction and a longing for inclusion (Kegan, 1982, 1994).

In order to demonstrate how individuals grapple with these forces as they move through the five orders of consciousness, Kegan (1994) made a distinction between subject and object. In this case, subject "refers to those elements of our knowing or organizing that we are identified with, tied to, fused with, or embedded in," whereas object "refers to those elements of our knowing or organizing that we can reflect on, handle, look at, be responsible for, relate to each other, take control of, internalize, assimilate, or otherwise act upon" (p. 32). For instance, college students who reason at the third order of consciousness – "cross-categorical thinking" – are able to construct their own point of view while also recognizing that others do the same (Love & Guthrie, 1999). This realization requires movement from understanding one's attitudes and values as part of oneself (i.e., subject) to something they encompass (i.e., object). Love and

Guthrie (1999) explained that prosocial behaviors require this type of knowing since it considers others as well as the self.

Although Kegan (1982, 1994) first defined meaning-making structures as a component of human development, it was Baxter Magolda (2001, 2009) who applied and expanded Kegan’s work to the study of college students. Baxter Magolda (2001) described a number of developmental tasks associated with college-going, including values exploration and path determination., including three primary questions students must ask themselves as they go through the process of getting to self-authorship: “How do I know?” “Who am I?” and “How do I want to construct relationships with others?” (p. 15). Each of these questions relates to the cognitive, intrapersonal, and interpersonal dimensions of meaning-making described by Kegan (1982).

Students begin their “journey toward self-authorship” with external meaning making (Baxter Magolda, 2001, p. 40). At this phase of “following formulas,” students define their knowledge and identity through external influences such as social norms and parental expectations. As students transition from this phase to internal meaning making, they are at a “crossroads” where they must “resolve the tension between what they wanted and what others wanted or expected” (Evans et al., 2010, p. 185). Jones and Abes (2013) described this phase as a time when “it is difficult to be confident in these growing internal beliefs, or even certain as to their precise nature, making acting on emerging internal ideas a struggle or an impossibility” (p. 100). Students eventually start to become the author of their own life, which is characterized by an ability to decide what they believe and defend these judgements in the face of conflicting external perspectives (Baxter Magolda, 2001). Finally, students are able to develop a “solidified and comprehensive system of belief” (Baxter Magolda, 2001, p. 155) which influences their sense of self and their relationships with others. Students at this concluding phase have complex meaning-making structures which allows them to understand external influences, yet trust their own attitudes when making decisions.

When applied to the RMMDI, meaning making capacity and self-authorship are drawn as a filter between the context and one's identity. Jones and Abes (2013) described how the permeability of this filter, which is based on the complexity of one's meaning-making capacity, determines how much the context influences one's identity. Students with more complex meaning-making capacities have narrow, less permeable screen openings; students with less complex meaning-making capacity have wide, more permeable opening. The permeability level of the filter matters since "contextual, external influences more easily move through a highly permeable filter (representing less complex meaning making), thereby having a stronger influence on a person's perceptions of identity than they would if the filter were less permeable (representing more complex meaning making)" (Jones & Abes, 2013, p. 104). It should be noted, however, that the filter is never impenetrable; identity will always be influenced by contextual factors no matter how complex a student's meaning-making capacity.

Applying the RMMDI to our understanding of bystander identity. To my knowledge, no research has examined how college students develop a bystander identity. However, the many components of the RMMDI make it a useful tool for understanding how college student bystanders develop their sense of self, which in turn influences their bystander disposition, tendencies, and behaviors. How one perceives their bystander attitudes and beliefs is represented by the core. These attitudes and beliefs can include factors also related to cognitive and moral reasoning, such as self-efficacy; they can also be related to non-cognitive processes, such as perceiving oneself as someone who intervenes when situations arise. Students with this stronger bystander “sense of self” should demonstrate more bystander tendencies and possibly increased future behaviors.

The interaction and salience of one’s social identities will also influence one’s bystander identity. For instance, white students may not notice racially-biased incidents in the same way students of color would because of the “prism of privilege and difference” (Jones & Abes, 2013, p. 86). Yet, white students have the ability and responsibility to intervene in these situations because of their white privilege. There is also evidence to suggest that women are more likely to intervene on behalf of other women at risk for sexual assault (Bennett, Banyard, & Edwards, 2015; Brown, Banyard, & Moynihan, 2014; Burn, 2009). Additionally, the ways that these social identities interact and intersect with one another also influences one’s bystander identity. As previously stated, individuals do not experience privilege and oppression based on their social identities in mutually exclusive ways. The salience of one’s intersecting identities also influences their bystander identity. Black women, who experience marginalization based on the intersections of race and gender, report engaging in bystander behaviors more often than white women (see Brown, Banyard, & Moynihan, 2014), indicating they may perceive their bystander identity differently than white women.

The environmental context is also an important factor in determining a student’s bystander identity. A campus culture which tolerates any form of bias toward underserved populations could hamper a student’s bystander identity; on the other hand, a campus culture which actively celebrates diversity and inclusion could advance one’s bystander identity. Additionally, the messages promoted by campus officials and educational programs provided by staff also influence a student’s bystander identity.

Finally, the RMMDI is more useful in this context than the original MMDI because it considers one's meaning making as a filter between context and identity. As bystanders observe and reflect on their surroundings, the complexity of their meaning-making capacities will influence their bystander identity. Students who demonstrate external meaning making will be more heavily swayed by their context in the development of their bystander identity than those with internal meaning making. For instance, new members of a fraternity or athletic team are more likely to follow the lead of other, more senior group members when determining their role as a bystander. As students develop in their self-authorship, they become less affected by their surrounding contexts. Thus, students who have intervention as part of their self-authored meaning-making structures will not be influenced by a context discouraging these behaviors.

Theoretical Perspectives Summary

The purpose of this section was to discuss the theoretical frameworks guiding bystander intervention disposition. The theories presented in this section demonstrate the complexity of the bystander decision-making processes. Those who witness instances of campus violence – no matter how seemingly ordinary – must use cognitive and moral reasoning as they notice the event and decide what actions to take. Several cost and reward factors contribute to this process, including the relationship to the victim and/or perpetrator; the perceived severity of the action and worthiness of the victim; the race, gender, and social class (and intersection of such identities) of the those involved and the possibility of shared group membership; and the presence of other bystanders. A bystander's level of empathic development can also influence the decision to intervene on behalf of a victim.

How college student bystanders perceive themselves as community members ready to intervene – their disposition – depends on a number of factors, including the environmental context, their cognitive and moral reasoning, and their identity as a bystander. The frameworks discussed in this section inform my study of collegiate bystander intervention disposition in a number of ways, and I believe using a combination of these perspectives is the best strategy to understanding this phenomenon. The socio-ecological framework developed by Bronfenbrenner (1979, 1993), and revised by Dahlberg and Krug (2002), guides the understanding of the various contextual factors underscoring bystander situations and behaviors. Additionally, the decision-making models articulated by Latané and Darley (1970) and refined by Pilivian et al. (1980) as well as the theories of bystander moral reasoning (Hoffman, 2000) illuminate the factors used by college student bystanders as they decide whether and how to intervene. Finally, the RMMDI (see Jones & Abes, 2013) highlights the role of identity development and meaning-making capacity in the bystander decision-making processes. These approaches, when taken in concert, contribute to a sophisticated understanding of collegiate bystander intervention disposition.

Chapter Conclusion: Conceptualizing Bystander Intervention Disposition

Within the context of higher education in the United States, violence unfortunately occurs in many ways for many people. One promising approach to ending many forms of campus violence is the training of students to act as prosocial bystanders. However, little is known about collegiate bystander intervention in contexts beyond sexual violence. This study attempts to expand our understanding of collegiate bystanders by considering the other forms of violence found on college and university campuses to measure bystander intervention disposition.

Bystander intervention disposition is related to a number of internal factors, including moral reasoning and empathy (Hoffman, 2000), self-authorship (Baxter Magolda, 2001; Kegan, 1982, 1994), social identities (Brown, Banyard, & Moynihan, 2014; Burn, 2009; Katz et al., 2017; Katz et al., 2018), and previous environmental contexts (Bourdieu, 1990; Bronfenbrenner, 1979, 1993; Carlo & Randall, 2001). For instance, collegiate bystanders with high moral reasoning and empathy are more likely to take responsibility for the assistance of others in harmful or negative situations. Bystanders with high self-authorship are also more likely to have higher intervention disposition because they can more easily filter out the social and cultural norms pressing them to not intervene. One's social identities and their perceived saliences, shaped by systems of privilege and oppression, additionally influence bystander intervention disposition; bystanders with many privileged identities may not notice certain situations as harmful to other marginalized identities or believe intervention behaviors are "worth it" when they do. On the other hand, bystanders with more than one intersecting marginalized identity may be more inclined to intervene on behalf of others in a variety of contexts due to their more complex understandings of oppression.

One's previous bystander intervention experiences and contexts, including how they were socialized to interact with similar and different others, can influence bystander intervention disposition. Those with high perception of community support or sense of belonging may be more inclined to intervene since they want to contribute positively to their community by stopping negative behavior. Additionally, the result of past bystander intervention behaviors can influence disposition to intervene. Bystanders who have witnessed others successfully intervene in the past, or who have successfully intervened themselves, are also more likely to intervene in the future. Finally, as bystanders mature and are exposed to more situations, they may also become more likely to intervene since they have seen the various consequences of standing by.

With these factors in mind, bystander intervention disposition can be conceptualized as a continuous latent construct ranging from low disposition to high. Collegiate bystanders with low intervention disposition will only intervene in uncomplicated or unchallenging situations (i.e., those that are easy and not personally dangerous). Students who exhibit high bystander intervention disposition, on the other hand, will intervene in any number of harmful situations, including those with perceived high costs.

Several aspects influence a situation's level of challenge. First, situations with higher or more perceived costs to bystander are tougher to endorse. Costs can include embarrassment, time and money, and social disapproval as well as physical harm (Pilivian et al., 1981), and are strongly influenced by community norms and culture and the bystander's relationships with the involved parties (i.e., victim(s), perpetrator(s), and/or other bystanders). For example, bystanders are more likely to find a harmful situation challenging if social norms do not support intervention behaviors (e.g., it causes a man to look weak in front of other male peers). Collegiate bystanders are also challenged by situations in which they do not know the victim; in these cases, it is more difficult for the bystander to empathize or feel a sense of responsibility to help. On the other hand, knowing the perpetrator or other bystanders can actually make a situation more difficult for bystander intervention. Adherence to social norms can be stronger when known others are present. Bystanders may additionally feel the cost of embarrassment if they attempt to intervene, but fail in front of those they know (i.e., audience inhibition). Finally, if the bystander knows the other bystanders and they fail to intervene, the bystander might find this situation difficult due to pluralistic ignorance or diffusion of responsibility.

Perceived cost to the victim also influences a situation's level of difficulty. When the harm to the victim is less clear, the situation becomes more challenging for the bystander. For instance, although microaggressive comments and actions are harmful in that they contribute to violence of marginalized populations, bystanders may not see them as harmful enough to warrant intervention. These types of situations also do not always have an explicit victim, making their level of harm more ambiguous to the bystander. The timing of the situation may also influence the victim's perceived cost. In many cases of campus sexual violence, collegiate bystanders witness the lead up to the incident and not the incident itself. It can be more challenging for bystanders to intervene if a situation is not currently harmful, but has the potential to become harmful.

A harmful situation's level of difficulty not only influences if a bystander will act, but how. It is harder for someone to act in the moment, especially if the victim's cost seems low and the bystander could experience high social costs or physical harm. In more challenging situations, the bystander may instead intervene by calling on other bystanders to help or indirectly intervening by informing an authority figure to stop the behavior. Bystanders can also intervene at a later time. If they know the victim and/or the perpetrator, it might be easier to for a bystander to reach out to them when the situation is not as intense or other bystanders are not present. Bystanders could also alert an authority figure of a situation after the fact to help prevent future similar behavior by the perpetrator.

When it comes to understanding and measuring bystander intervention disposition, scholars must consider an individual's personal characteristics as well as the circumstances of the situation. Some situations make intervention more challenging for bystanders, whereas other situations are easier for bystanders to interrupt. Collegiate bystanders with low intervention disposition are more likely to only intervene in easier situations with less costly behaviors. Bystanders with high intervention disposition, however, will intervene in more difficult situations using more costly behaviors. In the next chapter, I describe the design of an instrument intended to measure collegiate bystander intervention disposition across a variety of harmful situations found on college and university campuses.

Chapter 3: Methods

This chapter will focus on the research design and methodological processes used to create and test a new measure of college student bystander tendencies. Since this measure was developed following the qualitative work of Mayhew, Caldwell, and Goldman (2011) and the psychometric testing of Mayhew, Lo, Dahl, and Selznick (2018), the research design for this study begins with item generation and pilot testing on a sample of 1,939 students at one of three universities in the United States. Information regarding the sample and statistical methods used to test the validity and reliability of this instrument are also included. This chapter aims to accomplish the following purpose of the study: Investigate the validity of an instrument designed to measure collegiate bystander intervention disposition. Limitations of the study are also discussed.

Research Design and Methodologies

The purpose of study is to psychometrically test the reliability and validity an instrument developed to measure college student bystander disposition. Following DeVellis' (2018) and Wright and Stone's (1979) outlines of scale development (as discussed in the previous chapter), the methodological processes I plan to use to examine the reliability and validity of this instrument are outlined below. Before I begin with a full description of the methods used in this study, I would like to provide a description of my positionality as a researcher to give some context on how I approach this research.

Positionality

I would be remiss if I did not describe how my educational background in applied mathematics, student affairs administration, and higher education research, in addition to my professional experiences advising students in student affairs and higher education, have influenced my research approaches. As a mathematician, I cannot disregard the existence of the absolute truths. I assume that reality is physical and observable, even if some of us lack the language and awareness to know that it is there; that the purpose of research is used to explain and predict how people think, feel, and behave as best we can; and that there is a verifiable truth and if we measure something accurately, it should be considered true. These assumptions inform my decision to create a new instrument to measure bystander intervention as a latent trait, a decision that also indicates my epistemology is inherently postpositivist.

However, my work is still student development oriented; I believe knowledge and research can be broad and generalizable while also considering how multiple social groups might be impacted differently given systems of power and oppression that exist. Therefore, while a postpositivist worldview dominates my ways of thinking, I also apply critical approaches to inquiry, values, and knowledge accumulation to my scholarship. This criticality specifically grounds my understanding that individuals with different backgrounds, worldviews, and identities experience campus environments differently, and the holding of certain social identities provides some respondents with more privilege than others. This perspective has influenced my interest in bystander intervention generally and will further underpin my approaches to data analysis.

I also do not believe that researchers can be completely objective, even when using quantitative methods, since they always come to the methods with biases and preferences based on their backgrounds, worldviews, and identities. As such, I would like to name my personal characteristics that I believe are relevant to this study. I identify as a white, heterosexual, cisgender[4] woman from an upper-middle-class, well-educated family. Although I am not personally religious, I benefit from Christian privilege. Because of these many privileged social identities, I have also never found myself the target of identity-based harassment or bigotry. I also have not explicitly experienced sexual violence, except for the occasional microaggression. Additionally, as someone interested in bystander intervention, I find myself constantly aware of my surroundings and the possibility of situations which may call on me to act on behalf of others. In one such instance, I thought I saw someone drowning as I was walking on a trail along a river and called for emergency help. When it (thankfully) turned out to be nothing, I felt immense shame that I was the one who had called the police and caused a scene which included multiple fire trucks, a dive team, and a helicopter. Although the emergency personnel assured me I had done the right thing, I still question whether I would intervene in the same way in the future. These experiences provide me with some perspective and understanding as to how students may think about potentially harmful situations and the reasoning behind their behavior.

Phenomenon for Measurement: Collegiate Bystander Intervention Disposition

The phenomenon of interest for this instrument is collegiate bystander intervention disposition. I chose to focus on disposition rather than attitudes, beliefs, knowledge, or efficacy because disposition describes one’s inherent qualities of mind and character with regards to an inclination or tendency (Bourdieu, 1990; Weininger, 2002); in other words, disposition is a latent

[4] The term *cisgender* refers to “individuals who possess, from birth and into adulthood, the male or female reproductive organs (sex) typical of the social category of man or woman (gender) to which that individual was assigned at birth. Hence a cisgender person’s gender is on the same side as their birth-assigned sex, in contrast to which a transgender person’s gender is on the other side (trans-) of their birth-assigned sex” (Aultman, 2014, p. 61).

characteristic with implications for possible behaviors. Collegians with high bystander intervention disposition are more likely to intervene in challenging or difficult situations than students with low bystander intervention disposition. However, bystander intervention disposition is a latent psychological construct related to, but distinct from, bystander behaviors. There are some limitations to this point of view since prosocial bystander behaviors are what educators hope students will exhibit. However, empirically assessing student behavior oftentimes proves to be a challenging endeavor (see McMahon, 2015), if not an unethical one given the sensitivity of these situations. Therefore, I have decided to focus on measuring bystander intervention disposition in hopes that information regarding this construct will still enhance our understanding of bystander future behaviors (see Azjen, 1991, 2002).

Instrument Development

Once the theoretical models have been selected and the construct has been defined, the next step in instrument development is the creation of a large pool of items intended to measure the construct. The items developed for this study were written based on the qualitative research done by Mayhew et al. (2011) and the psychometric testing of Mayhew et al. (2018). The research of Mayhew et al. (2011) set out to define campus violence with the intent to create an instrument designed to assess the campus climate toward violence and safety. This scholarship led to the formation of scenario-based questions designed to measure student bystander intentions related to two forms of campus sexual violence: incapacitated sexual assault and domestic violence (see Mayhew et al., 2018). In Mayhew et al.'s (2018) quantitative work, student respondents were presented with a hypothetical situation in which another student needs the respondent to intervene. After reading the situation, respondents were then asked their likelihood to react using specified behaviors based on their relationship with the possible victim.

In the first scenario – a possible sexual assault at a party – a male student who has not been drinking very much is seen leaving a party with a female student who has clearly been drinking. After being presented with this scenario, respondents were then asked to respond to the behavioral items given a hypothetical relationship with the various characters – as a friend of the man, as a friend of the woman, and as someone who does not know either person very well. The behaviors for this situation included saying something, physically intervening, and getting other people at the party to support their intervention actions. Respondents were given the scenario and behaviors three distinct times, one for each possible relationship with the characters.

The second scenario, which addressed domestic violence, described an incident in which upstairs neighbors who are known by the respondent to be in a turbulent relationship are heard arguing. The scenario then instructs the respondents that they hear some noises that could indicate the situation has turned violent. Respondents were once again asked their likelihood to engage in certain behaviors, including saying something to the aggressive neighbor, saying something to the non-aggressive neighbor, getting other people to support an intervention, and calling the authorities (e.g., police, apartment manager) to intervene.

These questions have been used on several surveys related to collegiate residential environments and outcomes on the supposition that students residing in communities with strong ties (e.g., living learning programs or residential colleges) will be more likely to intervene than students in other residential environments (see Appendix A for the scenarios and items). Mayhew et al. (2018) examined the psychometric qualities of these questions on a sample of 2,846 students who responded to the 2015 and 2016 administrations of the Study of Living Learning Programs (SILLP) survey. The two scenarios with differing relational perspectives and possible behaviors resulted in 14 different items related to college student bystander intervention tendencies. After assuming a one-factor model, Mayhew et al. (2018) found these 14 items had high co-variability (Cronbach's alpha of 0.92). They also determined this approach acceptably modeled bystander intervention tendencies using confirmatory factor analysis, although they noted that further research was needed to fully determine its validity.

Since the purpose of this study is to examine college student bystander disposition across a number of possible scenarios, I developed an expanded set of bystander questions based on the scenario-based procedure utilized by Mayhew et al. (2018). These scenarios have been designed to address a number of other situations collegiate bystanders could encounter on campus, such as racial harassment in the residence hall, theft in the library, religious insensitivity by a faculty member in class, invasion of privacy during a sexual encounter, and a racially-biased post by a senior member of a student organization. In all of these situations, respondents have been provided with a specified relationship to either the victim, the perpetrator, or other bystanders. Additionally, they were asked to respond with the likelihood of reacting to the scenario with certain behaviors such as saying something, getting other people to intervene, and finding an authority figure; respondents could enact these behaviors in the moment or at a later time. See Appendix B for the full list of scenarios and response items.

A scenario-based approach was used to measure bystander intervention disposition since the contexts and environment in which students find themselves influence their perception of a situation and their potential behavior (Bronfenbrenner, 1979, 1993; Carlo & Randall, 2001; Jones & Abes, 2013). Using a scenario-based approach was also informed by Ajzen's (1991, 2002) theory of planned behavior (TPB). This theory posits that enacted behavior is ultimately a function of intentions that are shaped by three factors: attitude toward behaviors, subjective norms, and perceived behavioral control. Positioned within the context of this study, attitudes can be broadly considered as the extent to which students' perceive intervention as being favorable or unfavorable (i.e., cost-benefit analysis; Piliavin et al., 1981); subjective norms as the perceived social pressure to intervene, which can often be shaped by the campus context and climate; and perceived behavioral control as the perceived degree of difficulty associated with engaging in the behavior, which is often associated with both previous bystander experiences and anticipated impediments to intervention (see Azjen, 1991, p. 188).

I chose the theory of planned behavior for two primary reasons. First, given its emphasis on personal development and perceived locus of control, the TPB has gained recent support in literature associated with assessing effective bystander intervention behaviors outside of collegiate contexts (see Abbott & Cameron, 2014; Casey & Ohler, 2012; Stueve et al., 2006). Second, and perhaps more pragmatically, the TPB allows for the assessment of intentions to behave in scenarios that, in an ideal world, students would not actually encounter. In other words, while assessing actual bystander intervention behaviors would perhaps provide a better indicator of students' true disposition, such behaviors only come about in scenarios of distress where assessment of student actions would be impractical, if not unethical, as described above. Since disposition describes one's pre-reflexive thoughts, perceptions, and behaviors (Bourdieu, 1990; Weininger, 2002), it makes sense to measure this latent construct by asking students how they would act in a particular situation.

The scenarios and action items used in this instrument were written to cover the range of easy and difficult situations found on college campuses. Some are more difficult due to the ambiguity of the harmful incident, others are more difficult because of the relationships with the other parties present. For all of the situations, students are provided with possible behaviors that also range in level of difficulty; it is easier for a student to intervene at a later point or indirectly intervene by calling on someone else than it is to intervene directly in the moment. For instance, the situation with the faculty member stating religiously insensitive comments in class is most likely the hardest scenario on the survey. Not only are insensitive comments ambiguously harmful, but the faculty member has some power and authority over the students in the class. How students respond to the various behavior items in this challenging scenario additionally helps determine their disposition. Respondents who answer they would be likely to say something in the moment have higher bystander intervention disposition than those students who would not say something at all to their professor. Furthermore, these students who would say something in the moment have higher disposition than those students who intervene at a later time or by informing an authority figure.

At the other end of the contextual difficulty spectrum is the situation in which an intoxicated female friend of the respondent leaves a party with an unknown man. This scenario is less challenging for bystanders to interrupt because they are friends with the victim and should want to ensure her safety. This event is also commonly used by sexual violence training programs to encourage bystander behaviors at parties, so all collegians should encounter this situation with some intervention awareness, knowledge, and skills. As with the more challenging scenarios, the level of likelihood a student will enact the listed behaviors is an indication of their intervention disposition. Even in this situation, it is more difficult to intervene directly in the moment than at a later time or by intervening indirectly by telling an authority figure.

Instrument Administration

The scenarios and items used for this study were piloted on a group of college students who responded to the 2018 administration of the Assessment of Collegiate Environments and Outcomes (ACREO) survey. This project is an updated version of SILLP – the study on which

Mayhew et al. (2018) evaluated the original bystander items – and is designed to assess the relationship between a variety of on-campus experiences and behaviors and student outcomes. ACREO provides campus administrators with robust assessment and evaluation data while also producing data for researchers to continue and improve upon previous research on college residence life. As such, institutions self-select to participate as well as provide the sample of students to receive the survey. Each year, surveys are administered by the ACREO team under the auspices of the participating institution; institutional partners determine the survey administration dates and students receive the invitation email as if it were coming from a member of their university's staff.

As a project focused on measuring a variety of educational experiences and outcomes, the ACREO survey includes a number of items for constructs not explicitly related to bystander intervention disposition. Examples include campus sense of belonging, discussion of sociocultural issues with peers, and campus climate for underserved populations. In addition to the bystander scenario questions, ACREO also includes a set of items related to bystander knowledge of resources and bystander intention to report events of sexual violence and bullying. The items found on the ACREO survey were first piloted in 2015 with revisions made for the 2016 administration. The survey has not been changed since the 2016 revision except for the introduction of new items to measure additional constructs (e.g., innovation disposition, learning integration, financial literacy) in 2017 and 2018. The survey is administered annually during the spring semester at institutions who opt in.

Sample information. The most recent administration of this survey, which occurred during the spring of 2018, invited 12,893 college students at three public 4-year institutions across the United States to participate. Although each participating institution is classified as a public doctoral university with at least high research activity, they vary in their geographic location and institutional contexts. Two of the institutions are located in the Far West, while the third is in New England. Since environmental context influences how bystanders perceive and respond to acts of violence, detailed institutional information is provided in this section.

The first institution in the Far West is a land-grant university with very high research activity; it is the largest university in its state. At the time of survey administration, 47.1% of students identified as women, 24.8% as racial or ethnic minorities from the United States, and 11.7% as international students. Administrators at this institution founded an office of institutional diversity less than 5 years ago. Most of the students at Far West 1 are pursuing majors in the STEM disciplines.

The second institution in the Far West is its state's flagship university and also engages in very high research activity. During the 2017-2018 academic year, 53.3% of students identified as women, with 26.8% as students of color, and 11.8% as international students. This university has several centers and offices dedicated to diversity, equity, and inclusion, including a bias response team. Far West 2 is considered a very competitive institution, with top majors including the social sciences and psychology, journalism and communication, and business administration.

The New England university is the second-largest institution in its state and engages in high research activity. In the last 10 years, enrollment at this university has increased by 50%. At the time of administration, 36.6% of students identified as women, 33.8% as students of color, and 3.7% as international students. Although an office dedicated to institutional diversity does not currently exist on this campus, diversity and inclusion has been outlined as a pillar of the current strategic plan. Most students at this university pursue degrees in business administration, science or engineering, technology, and law enforcement.

The response rate for this study was 23.9%. Data cleaning, which removed participants with responses to less than 80% of the full survey, yielded a total sample of 1,939 students. The sample for this study included notably more students identifying as cisgender women (56.9%), with 39.9% identifying as cisgender men and 3.2% identifying with another gender identity. Students in the sample also primarily identify as heterosexual or straight (80.6%), with 9.6% identifying as bisexual, 6.1% identifying as queer or with another sexual orientation, and approximately 3.7% identifying as either gay or lesbian. Most of the respondents identified themselves as White/Caucasian (64.3%), with the remaining students representing Asian/Asian American, Native Hawaiian, or Pacific Islander (12.7%), multiracial or multiethnic (10.7%), Latino/a/x or Hispanic (6.6%), and Black/ African American (3.8%) backgrounds. The remaining 2.0% identified as "another race or ethnicity," including Native American and Middle Eastern. In terms of worldview/religious perspective – an identity characteristic increasingly found to be associated with numerous collegiate outcomes (see Bowman, Felix, & Ortis, 2014; Mayhew et al., 2017) – 41.5% of students in this sample consider themselves as holding a majority worldview (Christianity), 36.1% as non-religious, 10.7% hold another worldview, and 7.0% identify with a minority worldview such as Hinduism, Islam, or Judaism; 4.7% of the sample holds more than one worldview. The sample also includes a small number of international students (4.1%).

Additionally, 31.3% of participants self-identified as first-generation college students, meaning neither parent had attended college, including coursework toward an associate degree. The participants also represent a wide range of academic class years, with 60.5% of the sample in of their first year at college at the time of the survey administration, 18.9% in their second year, 11.3% in their third year, 6.5% in their fourth year, and 2.8% in their fifth year or more;

given that the survey was administered in the spring, all students had been on a college campus at least one semester at the time of response. The most common academic disciplines among these students are science, engineering, or mathematics (43.0%), social sciences or education (17.0%), business administration (13.2%), health professions (13.2%), and arts and humanities (7.4%); 6.2% of students did not indicate a major. Finally, 39.7% of students attended the public university in New England, with 36.7% attending one of the public universities in the Far West and 23.6% attending the other public university in the Far West. See Table 3.1 for student demographic information.

Table 3.1. Sample Characteristics (N=1,939)

Variable	Percent	N
Gender		
Cisgender man	39.9%	773
Cisgender woman	56.9%	1104
Genderqueer, transgender, or another gender identity	3.2%	62
Sexual Orientation		
Bisexual	9.6%	186
Gay	2.3%	45
Heterosexual/Straight	80.6%	1562
Lesbian	1.4%	27
Queer or another sexual orientation	6.1%	119
Race/Ethnicity		
Another race or ethnicity	2.0%	38
Asian/Asian American, Native Hawaiian, or Pacific Islander	12.7%	246
Black/African American	3.8%	74
Latino/a/x or Hispanic	6.6%	127
Multiracial or multiethnic	10.7%	208
White	64.3%	1246
Worldview/Religion		
Another worldview	10.7%	208
Nonreligious	36.1%	700
More than one worldview	4.7%	91
Worldview majority	41.5%	805
Worldview minority	7.0%	135

International Student		
No	95.9%	1859
Yes	4.1%	80
First-generation Student		
No	68.7%	1326
Yes	31.3%	605
Academic Class		
First year	60.5%	1168
Second year	18.9%	365
Third year	11.3%	218
Fourth year	6.5%	125
Fifth year or more	2.8%	54
Planned Academic Major		
Arts and humanities	7.4%	144
Business administration	13.2%	256
Health professions	13.2%	255
Science, engineering, or mathematics	43.0%	834
Social sciences or education	17.0%	329
No major selected	6.2%	121
Live on-campus		
No	3.3%	64
Yes	96.7%	1875
Institution		
Far West 1	36.7%	712
Far West 2	23.6%	457
New England	39.7%	770

The individual characteristics of the respondents are presented and considered in this study since they provide information related to identity and social perspective. Moradi and Grzanka (2017) recommended these demographic categories be reconceptualized as "dynamic social context variables" (p. 506) to help remedy several problematic practices currently used when these variables are used in quantitative research. For instance, reframing demographic characteristics in this way discourages atheoretical analysis of sociodemographic differences without considering similarities. Additionally, it encourages an equity-minded approach to theorizing how certain environments might influence outcomes across groups due to power dynamics instead of speculating what caused one group to differ from others.

How these individual characteristics intersect and overlap are also important to consider when analyzing the data. Although intersectionality should not be equated with “multiple identities” or “intersecting identities” (Moradi & Grzanka, 2017, p. 506), the various overlapping systems of power and oppression one experiences based on their multiple, intersecting identities may influence how individuals respond to the scenarios and items. For example, women may be more likely to intervene in the sexual violence scenarios than men, but how might depend on the intersections of race and gender; white women may be more likely to call on an authority figure such as a security or police officer in these situations than women of color, who may distrust these officers and the institutions they represent (Linder, 2018).

Administration of bystander intervention disposition items. Since the ACREO survey measures a number of constructs related to collegiate environments and outcomes, presenting all the possible scenarios and items related to this project would have significantly increased the length of the survey and decreased the completion rate. As such, the ACREO research team decided to randomly assign students to view and respond to certain new bystander disposition items. Every student who responded to the survey was given the opportunity to answer the items related to the scenarios found in the Mayhew et al. (2018) study (i.e., incapacitated sexual assault at a party and domestic violence). Then the respondents were randomly assigned to one of four groups using the survey software. Everyone in a single group saw the same set of scenarios and items, and the scenarios and items were mutually exclusive from group to group. Scenarios and items were designated to a certain group based on a number of factors, including type of incident, perceived ambiguity of the situation, and hypothetical relationship to characters in the scene. Table 3.2 describes the number of students in each group and the scenarios given.

Table 3.2. Bystander Groups (N=1,939)

Group	Percent	N	Scenarios & Items
Group A	23.5%	455	A1; C1; D3
Group B	24.3%	472	A2; C2; D4
Group C	25.1%	486	B1; D1; E1
Group D	25.3%	490	B2; D2; E2
No Branch	1.86%	36	

Methods of Analysis

Wolfe and Smith's (2007) Rasch validity framework was adopted to examine the psychometric functions of the collegiate bystander intervention disposition instrument. This approach was selected because it aligns specific Rasch analytic tools with each of the six components of Messick's (1995) unified concept of construct validity. The Rasch measurement technique, which was developed in the 1960s (see Rasch, 1960), has emerged as a preferred psychometric analysis framework for instrument development (Bond & Fox, 2015; Boone et al., 2014). This approach differs from other methods researchers employ to evaluate test and survey data (i.e., classical test theory) in that it holds to the principles of objective measurement (Boone et al., 2014; Thurstone, 1928). Objective measurement occurs when an instrument's calibration is independent of the objects used for calibration, and an object's measurement is independent of the instrument used to measure (Thurstone, 1928; Wright, 1967). Wright (1967) provided a helpful explanation of this concept:

> When a man says he is at the ninetieth percentile in math ability, we need to know in what group and on what test before we can make any sense of his statement. But when he

> says he is five feet eleven inches tall, do we ask to see his yardstick? We know yardsticks differ in color, temperature, compositions, weight - even size. Yet we assume they share a scale of length in a manner sufficiently independent of these secondary characteristics to give a measurement of five feet eleven inches objective meaning. We expect that another man of the same height will measure about the same five feet eleven even on a different yardstick. I may be at a different ability percentile in every group I compare myself with. But I am the same 175 pounds in all of them. (para. 8)

These conditions of objective measurement are what make it possible to generalize the measurement of an object beyond the instrument used and compare these measurements using similar, but not identical, instruments.

It should be mentioned that the sample independence aspect of objective measurement does not imply population independence of the instrument. Linacre (2018) noted this in his explanation of person-free measurement: “as much as is statistically possible, the item-difficulty estimates are independent of the particularly sample of persons from a homogeneous population that are used in the estimation” (p. 634). In other words, the calibration of the instrument should not change if a different sample from the same population is used to estimate those values. However, the calibrations are likely to change if a sample from a *different* population is used. We expect that an instrument designed to test math ability in 3rd graders will have different levels of difficulty if it were administered to high school seniors.

Rasch modeling is also less sensitive to missing data, which makes it an ideal analytic tool for evaluating an instrument to which respondents have not answered every item (Boone et al., 2014). As such, the research design used in this study, in which respondents did not view all the scenarios or items, will not influence the psychometric functioning of the Rasch model.

The Rasch model assumes the probability of a person endorsing an item is a function of their ability (or level of disposition) and the difficulty of the item (Boone et al., 2014). The model concludes that persons with more ability or disposition will endorse more difficult items, whereas persons with low ability or disposition will only endorse easier items. Moreover, people with a specific ability or disposition should only be able to correctly answer questions of a lower difficulty than their ability, and difficult items are less likely to be endorsed than easier items. This perspective, therefore, assumes that item difficulty and person ability can be measured on the same linear, continuous scale. When applied to rating scale data, such as those collected by Likert-type scales, Rasch analyses will linearize raw ordinal data into interval-level data by applying a natural log transformation to the matrix of item responses, as long as those items fit the Rasch model (Andrich, 1978; Bond & Fox, 2015; Boone et al., 2014; Ludlow & Haley, 1995). This process produces the new unit of measurement used to express person measures as well as item difficulties known as the logit (i.e., log-odds unit). Once data fits underlying Rasch models, it can be assumed estimates of collegiate bystander intervention disposition are linear, additive, interval-level, invariant, and hierarchical, and scores from the measure can be used in parametric statistics.

In order to use Rasch analysis to evaluate survey items, three assumptions must be met: construct unidimensionality, continuity of the latent construct, and item fit of the Rasch model (Boone et al., 2014). Construct unidimensionality refers to the alignment of the instrument's items with a single underlying latent construct. Sometimes this assumption is tested using exploratory factor analysis and parallel analysis prior to the Rasch analysis, but it can also be tested using principal component analysis (PCA) of the standardized residuals (Linacre, 2006). The second assumption, continuity of the latent construct, is often assumed given that an affective characteristic has an underlying continuum of internalization (Hopkins, 1998). The final assumption is confirmed through the process of fitting the data to the Rasch model (Boone et al., 2014).

The Rasch methods used to examine the validity of the collegiate bystander intervention disposition instrument were informed by Wolfe and Smith's (2007) conceptualization of Rasch validity evidence for Messick's (1995) unified concept of construct validity, which describes six aspects scholars should consider when "appraising the appropriateness, meaningfulness, and usefulness of score inferences" (Messick, 1995, p. 744). These six aspects include content, substantive, structural, generalizability, external, and consequential. However, since this instrument will not be used for standard setting or qualification of raters, the consequential aspect of validity is not examined. The activities used to evaluate the other five aspects of construct validity are described in detail below.

It should be noted that all methods of analysis were conducted using the Andrich-Wright Rating Scale Model (RSM) in the Winsteps (version 3.92.1) software. The RSM assumes that the distance between corresponding categories is the same across all the items. Rasch modeling also allows for the intervals between categories to differ across items with the Partial Credit Model (PCM). Although the RSM is generally favored when the same ordinal scale is used across items, as with the items in the collegiate bystander intervention disposition instrument, chi-squared test of difference was used to determine which model to use. Results indicate the PCM did not significantly improve the model fit and RSM should be used.

Additionally, all extreme (i.e., highest and lowest) respondent scores were excluded from the analysis. Boone et al. (2014) recommended this approach since the measurement error of the person measure is infinite for those who obtain the maximum or minimum score on an instrument, which can distort the analysis. This error occurs because when respondents get a perfect score, or the worst score possible, we are unable to measure how much more or less ability they have using the current instrument. As Boone et al. (2014) concluded, “if someone has topped out on an instrument… or if one has bottomed out…, these types of individuals do not provide useful data that help us understand how accurately the instrument is functioning” (p. 221). Exclusion of the extreme responses resulted in a final sample of 1,885 respondents used in the subsequent analysis. This sample represents 97.2% of the original sample.

Content Aspect of Construct Validity

The content aspect of construct validity refers to how well the items on the instrument represent the construct it is intended to measure (Messick, 1995). Wolfe and Smith (2007) identified the Rasch fit statistics as evidence supporting the content aspect of construct validity. Since the Rasch model conforms to the requirements of objective measurement, it does not adjust to the instrument, so items either “fit” the model, or they do not. Item fit refers to the how well items adhere to the expectations of the Rasch model (Boone et al., 2014). Items are determined to be “misfitting” when they do not perform as the Rasch model would predict.

The fit statistics provided by Winsteps describe how well items fit the Rasch model by examining the degree to which the item response patterns fit the model’s expectations. These statistics are reported as mean-square values (MNSQ), which are chi-square statistics divided by their degrees of freedom, and represent the amount of distortion the item contributes to the measurement of the instrument (Linacre, 2002). Although there are two types of fit – infit and outfit – the outfit statistics are more often used since they are usually easy to diagnose and remedy (Linacre, 2002). For instance, outfit is sensitive to outliers and will indicate items that do not fit the model due to lucky guesses and careless mistakes.

Items that contribute little distortion to the measurement system have an expected MNSQ value of 1.0 for both infit and outfit (Linacre, 2002). Linacre (2002) suggested survey items are productive for measurement when they exhibit outfit MNSQ between 0.5 and 1.5 logits; items degrade the measurement system when the exhibit outfit MNSQ greater than 2.0 logits and are less productive for measurement, but not degrading, when MNSQ is less than 0.5. Items with MNSQ values above 2.0 underfit the model and indicate unpredictability; MNSQ values below 0.5 suggest the data overfit the model and observations with these items are too predictable and possibly redundant. Linacre (2002) recommended that scholars evaluate items with high MNSQ over those with low MNSQ, with those items exhibiting high MNSQ being removed from the final instrument since they distort or degrade the measurement system. Boone et al. (2014) also suggests that well-fitting items exhibit a point-biserial correlation greater than 0.30 and close observed and expected point-biserial correlations (i.e., < 0.15). The fit statistics for the 83 items on the collegiate bystander intervention disposition instrument were evaluated using Table 10 in the Winsteps software.

Substantive Aspect of Construct Validity

The substantive aspect of construct validity emphasizes the role of theory when designing a new instrument (Messick, 1995). In other words, when the responses on an instrument align with the developer's intensions, it demonstrates substantive construct validity. Wolfe and Smith (2007) identified two indicators of substantive construct validity using Rasch methods. First, the rating scale used with the items should demonstrate monotonic functioning (i.e., average bystander intervention disposition increases with the values of the 5-point rating scale) and discernment (i.e., respondents can discriminate between the five response options) Second, the demonstrated difficulty of the items aligns with the difficulty predicted by theoretical conceptualizations.

To determine monotonic functioning of the rating scale, two tests were conducted using the Winsteps software. First, item polarity was assessed to check that the response categories have the correct order (i.e., persons with higher disposition select the greater response category). Item polarity is an issue when the average ability of the persons observed in a one response category is lower than the average ability of the persons in the next lower category (Linacre, 2018). Items with disordered categories were further analyzed using independent sample *t*-tests to see if the differences were significant. Category probability curves were additionally used to test the discrimination of the response scale. These curves demonstrate the region along the person measure for which a response category is most probable. Response category discrimination occurs when each probability curve has a region that is the most probable; if any curves are "buried" under other curves, there are too many response options and the number of scale categories should be reduced.

The level to which the Rasch-determined item difficulty aligns with theoretical conceptualizations is another indicator of substantive construct validity. In other words, items which are theoretically more challenging to endorse should have higher item difficulty values as measured by the Rasch model. The theoretical item hierarchy was assessed using the item difficulty values given by Winsteps Table 13. Items were then evaluated based on their associated scenario as well as the possible behavior for consistency with the theoretical considerations. It is hypothesized that the situations in which the respondent knows the victim will be easier to endorse intervention (see Batson et al., 2007; Burn, 2009; Katz et al., 2015). Furthermore, in those situations with high cost to the bystander, such as confronting an authority figure, or ambiguous cost to the victim(s), such as microaggressions, intervention will be more difficult to endorse. Finally, relationships with the perpetrator and other bystanders can influence the difficulty of a situation; knowing the other bystanders makes a situation more difficult when they do not intervene (see Brinkman et al., 2015; Latané & Darley, 1970), whereas knowing the perpetrator can make it more difficult since people hold more favorable views of their friends (see Ogletree & Archer, 2011; Tajfel & Turner, 1979). The hypothesized difficulty of the 16 scenarios in the instrument are provided in Table 3.3.

Table 3.3. Hypothesized Rank of Scenario Difficulty

Scenario Label	Description	Hypothesized Difficulty Rank
C2	A professor makes a religiously insensitive joke in class. Respondent has no friends in the class.	16
C1	A professor makes a religiously insensitive joke in class. Respondent has friends in the class.	15
D4	A floormate broadcasts a sexual encounter between another resident and a guest to other residents. Respondent is not friends with anyone.	14
S3	At a party with potential incapacitated sexual assault. Respondent is not friends with either the man nor the woman.	13
B1	Someone leaves their belongings unattended in the library. Respondent is friends with a person seen rummaging through the unattended belongings.	12
S4	Respondent hears neighbors arguing and it starts to sound physically violent.	11
D1	A floormate broadcasts a sexual encounter between another resident and a guest to other residents. Respondent is friends with the floormate.	10
D3	A floormate broadcasts a sexual encounter between another resident and a guest to other residents. Respondent is friends with the other bystanders.	9
B2	Someone leaves their belongings unattended in the library. Respondent is not friends with a person seen rummaging through the unattended belongings.	8
E1	A leader of a student organization posts a video of them saying racial slurs online. Respondent is a member of the organization and knows the student leader.	7

A2	Floormates write a racial slur directed at a student in the residence hall. Respondent is not friends with the victim.	6
S1	At a party with potential incapacitated sexual assault. Respondent is friends with the man.	5
D2	A floormate broadcasts a sexual encounter between another resident and a guest to other residents. Respondent is friends with the resident in the broadcast.	4
A1	Floormates write a racial slur directed at a student in the residence hall. Respondent is friends with the victim.	3
E2	A leader of a student organization posts a video of them saying racial slurs online. Respondent is also a leader of the organization and knows the student leader.	2
S2	At a party with potential incapacitated sexual assault. Respondent is friends with the woman.	1

In terms of the hypothesized difficulty of the behaviors, items were grouped into one of 11 possible actions based on the timing of the intervention behavior and the respondent's possible relationships with the victim, perpetrator, and other bystanders. These groupings cut across all the scenarios. It is hypothesized that saying something at the time to an unknown victim is the most difficult behavior for bystanders to take since the victim's costs are a more ambiguous (Piliavin et al., 1981). On the other hand, saying something at the time to a known victim is the easiest behavior since bystander's will intervene on behalf of their friends (see Batson et al., 2007; Burn, 2009; Katz et al., 2015). It is also difficult for bystanders to follow up with unknown victims or perpetrators due to the lack of relationship with these parties, whereas saying something later to known victims or perpetrators is easier; as time passes, bystanders can recognize the harm or feel more comfortable in engaging their friends in conversation after the fact (Carlo & Randall, 2011; Piliavin et al., 1981). Table 3.4 presents the hypothesized ranking in behavior difficulty.

Table 3.4. Hypothesized Rank of Behavior Difficulty

Items	Behavior Description	Hypothesized Difficulty Rank
s1_b, s3_b, s4_b, a2_b, d1_b, d3_b, d4_b	Say something at the time to an unknown victim	11
s4_d, a2_d, b1_c, b2_b, d1_d, d3_d, d4_d	Say something later to an unknown victim	10
s4_c, a1_c, a2_c, c1_b, c2_b, e2_b, d2_c, d3_c, d4_c	Say something later to an unknown perpetrator	9
s2_a, s3_a, s4_a, a1_a, a2_a, b2_a, c1_a, c2_a, e2_a, d2_a, d3_a, d4_a	Say something at the time to an unknown perpetrator	8
s1_e, s2_e, s3_d, s4_f, a1_f, a2_f, b1_e, b2_d, c1_d, c2_d, d1_f, d2_f, d3_f, d4_f, e1_d, e2_d	Call an authority figure	7
c2_c, d1_e, d2_e, d4_e	Get unknown others to support	6
s1_a, b1_a, e1_a, d1_a	Say something at the time to a known perpetrator	5
s1_d, s2_d, s3_c, s4_e, a1_e, a2_e, b1_d, b2_c, c1_c, d3_e, e1_c, e2_c	Get known others to support	4
s1_c, b1_b, e1_b, d1_c	Say something later to a known perpetrator	3
s2_c, a1_d, d2_d	Say something later to a known victim	2
s2_b, a1_b, d2_b	Say something at the time to a known victim	1

Structural Aspect of Construct Validity

The structural aspect of construct validity assesses the consistency of an instrument's internal scoring structure with the content domain (Messick, 1995). That is, "the relevance and representativeness of the test content in relation to the content of the domain about which inferences are to be drawn or predictions made" (Messick, 1989). To determine if an instrument holds to this aspect of validity, Wolfe and Smith (2007) recommended using principle component analysis (PCA) on the standardized residuals from the Rasch scaling of the data to test dimensionality when the intended structure is unidimensional. Linacre (n.d.) explained that this method:

> looks for patterns in the part of the data that does not accord with the Rasch measures. This is the "unexpected" part of the data. We are looking to see if groups of items share the same patterns of unexpectedness. If they do, then those items probably also share a substantive attribute in common, which we call a "secondary dimension." (para. 2)

The PCA approach was used since this instrument was designed to measure bystander intervention disposition as a single, unidimensional construct. Linacre (n.d.) suggested that a secondary dimension could be present if the first contrast of the PCA exhibits an eigenvalue greater than two, the minimum number of items needed for a dimension to be considered. If the eigenvalue of the first contrast is greater than two, the content of the items at the top and bottom on the contrast plot should be analyzed to determine what the items have in common that contrasts with the other items. Table 23 from the Winsteps software provided the output from the PCA on the standardized residuals.

Generalizability Aspect of Construct Validity

The generalizability aspect of construct validity concerns the degree to which an instrument is limited in its ability to measure the content domain by the items or sample used (Messick, 1995). Instruments demonstrating the generalizability aspect of construct validity maintain their integrity across various contexts and respondents. Three approaches were used to examine this aspect of validity. First, the profile of misfitting persons were examined using chi-square tests of independence to determine if a relationship exists between certain demographic characteristics and misfit to the Rasch model. A disproportionate number of misfitting responses for specific demographic groups may suggest there was some sort of "issue" influencing the manner in which the instrument is measuring (Boone et al., 2014, p. 163).

Person fit is similar to item fit, except that the focus is on the quality of the responses and not item performance. As an objective measurement technique, the Rasch model does not adjust to the responses for an instrument. Therefore, just as with the items, respondents either fit the model or not. Responses fit the model when they perform as the Rash model expects. Person misfit occurs when a respondent's actual pattern of responses diverges from that predicted by the model (Boone et al., 2014). As Boone et al. (2014) explained:

> Person fit looks at how a person answered all the items on a survey or test, but those answers are reviewed in light of the person's measure, which is computed using all of the respondent's answers compared to the difficulty level of the items. (p. 160)

Misfit can therefore either occur because responses are too predictable (i.e., overfitting) or unpredictable (i.e., underfitting) based on the predicted difficulty of the items.

As with item fit analysis, person fit indices are reported as infit and outfit mean-square values (MNSQ) as well as z-standardized values (ZSTD), calculated as the probability of the MNSQ value occurring by chance when the data fit the Rasch model. Person outfit ZSTD values greater than 3.0 indicate responses that underfit the model and vary unexpectedly from perfect fit; person outfit ZSTD values less than -3.0 overfit suggest the responses overfit the model and are too predictable (Linacre, 2002). The fit statistics for the 1,885 non-extreme respondents to the collegiate bystander intervention disposition instrument were noted using Winsteps Table 6; respondents demonstrating outfit ZSTD greater than 3.0 were flagged as underfitting and those with outfit ZSTD less than -3.0 as overfitting; respondents were also classified as generally misfitting if they either underfit or overfit the model. Chi-square tests of independence were conducted for gender, sexual orientation, race, worldview/religion, nationality, education generation status, academic major, and academic class year to look for any disproportionate numbers of misfitting, underfitting, and overfitting persons by demographic group. Chi-square tests of independence were also used to determine if a respondent's institution or randomly assigned group of scenarios could have influenced misfit.

Second, I observed Wolfe and Smith (2007) suggestion to use item difficulty invariance to examine the generalizability aspect of validity. Invariance occurs when an item's difficulty does not vary as a function of the sample (or subsample) used to derive the estimates (see Boone et al., 2014; Wolfe & Smith, 2007). Items that fail to demonstrate invariance exhibit differential item functioning (DIF) in that "the item defines a trait in a different manner when its performance is compared across two or more groups of respondents" (Boone et al., 2014, p. 275). When DIF is detected, "we may suspect that the item's content may provide an advantage to respondents within one of the groups or may disadvantage respondents in the other group" (Wolfe & Smith, 2007, p. 216). In other words, the item may be biased.

The presence of item DIF was assessed across respondent university (Far West 1, Far West 2, and New England), gender identity (cisgender man, cisgender woman, and genderqueer or another gender), and race or ethnicity (African American/Black, Asian/Asian American or Native Hawaiian or Pacific Islander, Latino/a/x or Hispanic, more than one race or ethnicity, Native American or another race or ethnicity, and white), using the recommended probability of contrast between item difficulty of $\alpha = 0.05$ (Linacre, 2013). The use of multiple comparisons was accounted for using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), which adjusts the critical value for each individual comparison to maintain an overall Type I error rate of 0.05. As such, the adjusted *p*-value for the DIF analysis across university and gender identity was 0.0002 and the adjusted *p*-value for the DIF analysis across race or ethnicity was 0.0001. For any items with statistically significant DIF, the "effect size" of the potential DIF was examined using the DIF contrast value to determine if the DIF is meaningful (Boone et al., 2014, p. 282). A DIF contrast less than -0.64 or greater than 0.64 indicates moderate to large DIF (Linacre, 2018) and should be flagged as potentially biased.

Finally, Wolfe and Smith (2007) recommended using reliability analysis to examine the generalizability aspect of construct validity, as these estimates provide empirical evidence that an instrument measures consistently over time and people (Boone et al., 2014). Reliability of the person measures was calculated to examine internal consistency of the instrument. Linacre (2018) noted that "person reliability" can be interpreted similarly to Cronbach's alpha, meaning that values closer to 1 indicate more internal consistency in the instrument. Winsteps Table 3 provides two types of person reliability estimates: the model reliability, which is the upper limit, and the real reliability, which is the lower limit (Boone et al., 2014). Both estimates were considered for this study.

External Aspect of Construct Validity

The external aspect of construct validity tests the extent to which an instrument's scores reflect the expected relations with other measures and behaviors based on the theory of the construct being measured (Messick, 1995). "Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures – or the lack thereof – are consistent with that meaning" (Messick, 1995, p. 746). Wolfe and Smith (2007) recommended using group comparisons as well as person-item maps to explore the external aspect of validity.

Group comparisons by demographic categories (gender, sexual orientation, race or ethnicity, worldview/religion, academic major, and academic class year) were used to document expected between-group differences on the measure. For each demographic group, a one-way between-subjects ANOVA was conducted to compare the average bystander intervention disposition score across the different demographic characteristics. For those ANOVA results with significant differences, post hoc analyses using the Scheffé test post hoc criterion for significance was used to test pairwise differences.

It is hypothesized that cisgender women will exhibit higher bystander intervention disposition than cisgender men (see Bennett, Banyard, & Edwards, 2015; Brown, Banyard, & Moynihan, 2014; Burn, 2009; Tjaden & Thoennes, 2000). Additionally, Black students are hypothesized to demonstrate higher bystander intervention disposition than white students (see Brown, Banyard, & Moynihan, 2014; Kunstman & Plant, 2008). Research on bystander intervention attitudes and behaviors is limited for other types of social identities, so it is difficult to hypothesize if minoritized and underrepresented groups will also have higher bystander intervention disposition, on average, than their more privileged peers.

Furthermore, students in humanities and arts majors will exhibit higher bystander intervention disposition than those in other majors since these majors typically use smaller, discussion-based courses and focus more on critical thinking, integrative learning, and issues of social justice (see Abbott & Cameron, 2014; Banyard et al., 2016; Brinkman et al., 2015). Students who have been on campus longer are also hypothesized to exhibit higher bystander intervention disposition since they have stronger sense of belonging and more developed self-authorship (see Banyard et al., 2016; Carlo & Randall, 2001). Finally, one-way between subjects

ANOVA tests were conducted for institution and randomly assigned group of scenarios. It is hypothesized that these categories will not have an effect on bystander intervention disposition.

Person-item maps were also used to examine the external aspect of validity. These maps are useful tools in determining how item level of difficulty aligns with respondent measurement level. Since the Rasch model transforms person scores into the same scale used to determine item difficulty, these two traits can be mapped together along the trait continuum (Boone et al., 2014). Person-item maps use the unidimensional logit scale found in Rasch measurement and present items on the right with respondents' abilities shown on the left. The easiest items are at the bottom of the map, whereas the more difficult items are at the top of the map; respondents with the lowest ability are located at the bottom of the map, and those with the highest ability are located at the top of the map. The mean item measurement is located at zero on the logit scale. Table 1 from Winsteps provided the person-item maps.

Limitations

Before concluding this chapter, I would like to recognize several of the limitations that should be considered in concert with the findings. The first is the sample[5] used to test the validity of the collegiate bystander intervention disposition instrument. Although students at three different universities received and responded to the survey, this sample is not nationally representative, so inferences about the population should be limited to similar students attending similar institutions in similar locations. Since context matters to understanding violence and how people perceive it (Dahlberg & Krug, 2002), these respondents may have had additional

[5] Although the Rasch model is sample independent, it is not population independent, and the sample determines the population. As such, the sample is important to note here since it provides information about who is in the population.

experiences that could influence their bystander intervention disposition. For instance, although the institutions included in this study did not matter in terms of item invariance or person misfit, respondents at one institution had significantly higher bystander intervention disposition than respondents at the other two. This finding emphasizes the role that context plays in determining bystander intervention disposition. For instance, the institution with the highest average bystander intervention disposition scores has the most extensive policies and practices for diversity and inclusion. Additionally, the three institutions represented in this study are in more traditionally liberal regions of the United States. Different findings could have occurred had the dataset included respondents from more conservative areas. As such, further research at other institutions across the United States should be considered to account for these environmental factors.

Another limitation is the lack of related measures on the ACREO survey. In addition to group comparisons, Wolfe and Smith (2007) recommend correlating the scores from the instrument with those of associated and unassociated measures to test for external validity. If the scores from the instrument of interest highly correlate with related measures, then these is more evidence to support external validity. Since the ACREO survey did not include these types of measures, this study was unable to use correlations to examine this aspect of construct validity. Future testing of this instrument should also include measures possibly related to bystander intervention disposition, such as social desirability (Crowne & Marlowe, 1960), cultural intelligence (Earley & Ang, 2003; Van Dyne et al., 2012), personality (Gosling et al., 2003), self-authorship (Creamer, Baxter Magolda, & Yue, 2010; Pizzolato, 2007), and moral reasoning (Ray, 2007). It should also ask respondents about their past experiences as bystanders, including in which situations they have intervened, as the construct should be related to bystander behavior. Additional testing with these measures would strengthen the case for the external aspect of construct validity.

A third limitation concerns one of primary assumptions underpinning the way the scenarios and items were constructed on the instrument. Embedded in the approach used is the presumption that if respondents did not recognize these scenarios as worthy of intervention, it meant they had low intervention disposition. In other words, those respondents who did not think a situation necessitated intervention would select “very unlikely” to the behavior items, which would suggest low intervention disposition. Although noticing an event and deciding it needs intervention are the first two steps in Latané and Darley’s (1970) Decision Model of Helping, there is a fundamental difference between not knowing a situation requires intervention and actively deciding not to intervene when the need is recognized. At this time, however, the instrument is unable to detect which one of these reasons influences the low bystander intervention scores.

Similarly, this instrument is limited by its conceptualization of intervention disposition. Bystanders to harmful situations have three response options – act to alleviate the harm, do nothing, or contribute to the hurtful behavior (Banyard, 2015) – yet this instrument does not consider the third option. Right now, all of the action items are prosocial in nature, suggesting that low likelihood to intervene means a tendency to do nothing. However, a low likelihood to intervene may also convey a willingness to contribute to the harm. Future iterations of this instrument may consider a continuum which ranges from harm contribution to walking away to helping the victim.

The final limitation involves the process by which the instrument was created. The scenarios and items were developed following the qualitative work of Mayhew et al. (2011) and the subsequent preliminary psychometric testing of the sexual and partner violence scenarios by Mayhew et al. (2018), which provided an adequate foundation for further exploration of bystander intervention disposition across other contexts and types of violence. However, the new situations developed for this study did not undergo cognitive interviewing and expert evaluation, as is recommended by Wolfe and Smith (2007) and others. Additionally, the theoretical ranking of the scenarios and bystander behaviors was completed by the author using only the literature as a guideline. These rankings would have benefitted from expert review by professionals who study and work with college and university students to prevent other forms of campus violence. Future refinement of this instrument should include cognitive interviewing as well as expert review to increase the evidence supporting its validity.

Chapter Conclusion

The purpose of this chapter was to describe the analytic methods used to validate a new instrument for measuring collegiate bystander intervention disposition. Using Wolfe and Smith's (2007) framework as a guideline, Rasch methodologies were used to examine five of Messick's (1995) six aspects of construct validity. In the next chapter, I present the results of the analysis in more detail.

Chapter 4: Results

This chapter provides the results of the methodological approaches outlined in Chapter 3. Since Wolfe and Smith's (2007) conceptualization of Rasch validity evidence for Messick's (1995) unified concept of construct validity informed the methods used in this study, the results will be presented using the same framework. All analyses were conducted using the Andrich-Wright Rating Scale Model (RSM) with all extreme respondent scores excluded from the analysis[6] (97.2% of original sample). Table 4.1 presents the global fit statistics for the RSM versus Partial Credit Model (PCM). The chi-square test of difference indicated that the PCM did not significantly improve the model fit, so the RSM should be used ($\chi^2_{160939-159406} = \chi^2_{1533} = 161141.2763 - 159585.5397 = 1555.7366, p = 0.337$).

Table 4.1. Global fit statistics for Rating Scale Model (RSM) versus Partial Credit Model (PCM)

Model	Log-likelihood chi-squared	d.f.
RSM	161141.2763	160939
PCM	159585.5397	159406

Content Aspect of Construct Validity

Evaluation of item fit to the Rasch model was conducted to examine the content aspect of construct validity (Wolfe & Smith, 2007). Item fit describes the extent to which the items adhere

[6] As noted in the previous chapter, Boone et al. (2014) recommended excluding respondents with extreme scores from subsequent analysis. Extreme scores occur when respondents obtain the maximum or minimum score on an instrument. These scores distort the analysis since the measurement error of a person with an extreme score in infinite.

to the predictions of the Rasch model (Boone et al., 2014). Table 4.2 presents the item infit, outfit, and point-biserial statistics. These statistics are presented in descending order by outfit MNSQ. Results indicate that the outfit MNSQ values for all items fall below the 2.0 logit threshold for productive measurement (Linacre; 2002). Additionally, all items but one (a2_g) exhibit a point-biserial correlation greater than 0.30 and five items (e2_a, a2_g, c1_a, c2_a, and s4_a) demonstrate differences in observed and expected point-biserial correlations greater than 0.15, the upper limit recommended by Boone et al. (2014).

Substantive Aspect of Construct Validity

The two indicators suggested by Wolfe and Smith (2007) – rating scale functioning and theoretical predictions – were used to examine the substantive aspect of construct validity. All items but five exhibited monotonic functioning for the rating scale categories. Table 4.3 presents the item polarity statistics for these five items that demonstrated response category mis-order based on mean ability for respondents selecting that category. Independent sample *t*-tests, however, indicate that the mis-order exhibited by these items is not statistically significant. In other words, although the average ability may indicate mis-order for this response categories, the differences in the average ability between the disordered categories is negligible. Table 4.4 presents the summary of the response category structure. None of the average response categories are mis-ordered, indicating that, on average, each level of the rating scale is substantively associated with a higher level of bystander intervention disposition. Additionally, the category probability curves suggest respondents are able to discriminate between the response categories currently used in this study (see Figure 4.1).

Table 4.2. Item statistics: Misfit order (by outfit MNSQ)

Item Label	Score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Observed Point-Biserial Correlation	Expected Point-Biserial Correlation	Estimated Discrimination
e2_a	1316	463	-0.215	0.05	1.76	9.41	1.94	9.79	0.34	0.53*	0.27
a2_g	1469	452	-0.803	0.06	1.61	6.79	1.87	7.59	0.28*	0.45*	0.56
e2_b	1518	468	-0.786	0.06	1.39	4.65	1.73	6.76	0.38	0.46	0.80
s1_c	5682	1877	-0.469	0.03	1.21	5.81	1.65	9.90	0.35	0.50	0.75
a1_g	1494	438	-1.099	0.07	1.44	4.65	1.53	4.53	0.34	0.45	0.78
s2_c	6123	1876	-0.828	0.03	1.26	6.25	1.48	9.36	0.38	0.46	0.86
a1_b	1236	439	-0.188	0.05	1.19	2.70	1.42	4.83	0.42	0.53	0.76
c1_a	593	436	1.326	0.05	1.19	2.81	1.40	5.13	0.39	0.59*	0.57
e2_d	1496	468	-0.708	0.06	1.41	4.98	1.37	3.83	0.40	0.48	0.79
b1_b	1448	464	-0.599	0.06	1.34	4.33	1.35	3.87	0.42	0.49	0.79
d1_f	1043	462	0.410	0.05	1.11	1.75	1.35	4.88	0.47	0.57	0.75
e1_a	1173	456	0.084	0.05	1.32	4.61	1.34	4.41	0.49	0.55	0.70
e1_d	1399	459	-0.496	0.06	1.35	4.48	1.33	3.73	0.42	0.50	0.79
b1_a	1570	465	-1.031	0.07	1.30	3.42	1.33	3.14	0.37	0.45	0.87
c2_a	590	452	1.385	0.05	1.25	3.69	1.32	4.18	0.41	0.61*	0.58
s4_a	3406	1843	0.826	0.02	1.16	5.30	1.31	9.00	0.42	0.60*	0.56
s2_a	5250	1874	-0.180	0.03	1.20	5.79	1.28	6.87	0.47	0.53	0.83
b2_b	1390	465	-0.406	0.05	1.20	2.72	1.27	3.14	0.46	0.51	0.89
d2_f	1110	460	0.266	0.05	1.21	3.27	1.27	3.69	0.49	0.58	0.74
d3_f	1032	432	0.294	0.05	1.18	2.69	1.26	3.55	0.45	0.56	0.70
b1_e	1026	464	0.459	0.05	1.18	2.86	1.26	3.80	0.43	0.58	0.60

Item Label	Score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Observed Point-Biserial Correlation	Expected Point-Biserial Correlation	Estimated Discrimination
a1_d	1260	438	-0.260	0.05	1.07	1.04	1.25	3.00	0.50	0.52	0.99
c1_b	804	437	0.833	0.05	1.18	2.90	1.24	3.37	0.48	0.59	0.71
e2_c	1409	468	-0.434	0.05	1.24	3.25	1.23	2.72	0.45	0.51	0.85
c2_b	807	453	0.890	0.05	1.04	0.63	1.16	2.46	0.52	0.60	0.89
c1_d	942	434	0.516	0.05	1.13	2.12	1.15	2.20	0.47	0.57	0.78
c2_d	972	454	0.531	0.05	1.07	1.15	1.13	1.92	0.53	0.58	0.90
a1_f	1226	439	-0.161	0.05	1.13	1.85	1.13	1.62	0.48	0.53	0.90
e1_b	1364	460	-0.386	0.05	1.19	2.67	1.13	1.58	0.50	0.51	0.94
d3_a	1150	434	0.010	0.05	1.05	0.78	1.13	1.67	0.50	0.54	0.95
s4_f	5421	1850	-0.335	0.03	1.07	2.14	1.11	2.81	0.45	0.52	0.92
a2_f	1296	457	-0.220	0.05	1.15	2.25	1.10	1.36	0.50	0.52	0.92
d3_b	1019	431	0.312	0.05	1.00	-0.04	1.10	1.40	0.52	0.56	0.96
s4_c	3719	1844	0.656	0.02	1.03	0.92	1.10	2.98	0.49	0.59	0.84
c1_c	904	435	0.604	0.05	1.05	0.88	1.10	1.43	0.49	0.58	0.87
d4_f	1093	443	0.196	0.05	1.13	2.08	1.09	1.33	0.54	0.56	0.91
a1_a	1272	440	-0.279	0.05	1.15	2.08	1.09	1.12	0.52	0.52	0.97
b1_d	1137	462	0.208	0.05	1.07	1.20	1.09	1.26	0.51	0.56	0.90
d2_d	1464	463	-0.653	0.06	1.19	2.43	1.08	0.87	0.46	0.48	0.97
d4_d	1061	443	0.273	0.05	0.95	-0.77	1.07	1.08	0.57	0.56	1.12
a2_a	1176	457	0.084	0.05	1.06	1.03	1.05	0.78	0.53	0.55	0.97
s2_e	4175	1863	0.430	0.02	1.01	0.31	1.05	1.55	0.52	0.58	0.93
b2_c	1203	467	0.091	0.05	1.03	0.51	1.05	0.67	0.49	0.56	0.93

									Observed	Expected	
									Point-	Point-	
Item				Model	Infit	Infit	Outfit	Outfit	Biserial	Biserial	Estimated
Label	Score	Count	Measure	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	Correlation	Correlation	Discrimination
b1_c	1407	464	-0.472	0.06	1.09	1.25	1.04	0.58	0.49	0.51	0.98
c2_c	890	453	0.705	0.05	1.03	0.48	1.04	0.70	0.55	0.59	0.98
d2_b	1459	464	-0.623	0.06	1.16	2.12	1.04	0.44	0.49	0.48	1.00
e1_c	1296	460	-0.196	0.05	1.02	0.33	1.03	0.36	0.54	0.53	1.05
s2_b	6594	1873	-1.347	0.04	1.10	2.13	1.02	0.36	0.44	0.41	1.04
d1_b	1081	463	0.336	0.05	0.98	-0.36	1.02	0.30	0.55	0.57	1.02
s1_e	3558	1877	0.771	0.02	0.92	-2.88	1.01	0.37	0.51	0.59	0.93
d4_a	1094	443	0.192	0.05	1.03	0.49	1.01	0.16	0.58	0.56	1.07
b2_d	1120	465	0.264	0.05	1.02	0.27	1.00	-0.01	0.52	0.57	0.97
d2_a	1344	463	-0.288	0.05	1.03	0.54	0.98	-0.30	0.53	0.52	1.07
d1_a	1352	463	-0.324	0.05	1.03	0.45	0.98	-0.30	0.51	0.52	1.04
s4_b	4427	1848	0.274	0.02	0.93	-2.23	0.97	-0.92	0.54	0.57	1.08
s3_d	3761	1851	0.641	0.02	0.93	-2.54	0.97	-1.09	0.56	0.59	1.07
a2_b	1138	457	0.170	0.05	0.89	-1.78	0.96	-0.54	0.55	0.55	1.12
s1_b	5072	1879	-0.060	0.03	0.92	-2.45	0.96	-1.11	0.53	0.54	1.09
s4_d	4843	1848	0.034	0.02	0.94	-1.81	0.95	-1.31	0.54	0.55	1.10
d1_d	1150	461	0.167	0.05	0.93	-1.11	0.94	-0.88	0.56	0.56	1.11
b2_a	1502	465	-0.772	0.06	1.03	0.48	0.94	-0.69	0.49	0.48	1.05
d2_c	1260	462	-0.080	0.05	0.95	-0.81	0.92	-1.13	0.58	0.54	1.15
s1_d	5203	1877	-0.143	0.03	0.88	-3.88	0.92	-2.30	0.54	0.53	1.16
d4_b	978	443	0.465	0.05	0.90	-1.67	0.89	-1.65	0.61	0.58	1.20
d3_d	1120	431	0.074	0.05	0.89	-1.83	0.89	-1.56	0.56	0.54	1.19

									Observed	Expected	
									Point-	Point-	
Item				Model	Infit	Infit	Outfit	Outfit	Biserial	Biserial	Estimated
Label	Score	Count	Measure	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	Correlation	Correlation	Discrimination
s1_a	5998	1877	-0.718	0.03	0.91	-2.57	0.88	-2.96	0.50	0.48	1.13
a1_c	1174	437	-0.036	0.05	0.90	-1.61	0.87	-1.78	0.59	0.54	1.21
d3_c	1077	433	0.184	0.05	0.84	-2.59	0.87	-1.91	0.59	0.55	1.21
a2_e	1245	455	-0.102	0.05	0.89	-1.76	0.86	-1.95	0.58	0.53	1.21
a2_d	1196	454	0.012	0.05	0.88	-1.97	0.86	-2.05	0.58	0.54	1.22
a2_c	1111	456	0.226	0.05	0.87	-2.21	0.86	-2.22	0.60	0.56	1.21
d4_c	1040	443	0.323	0.05	0.88	-2.03	0.84	-2.41	0.64	0.57	1.29
s3_a	4012	1861	0.515	0.02	0.82	-6.46	0.84	-5.13	0.60	0.58	1.22
s2_d	5713	1872	-0.503	0.03	0.89	-3.31	0.84	-4.10	0.55	0.50	1.17
d3_e	1099	435	0.146	0.05	0.85	-2.50	0.84	-2.35	0.60	0.55	1.24
s4_e	5056	1850	-0.092	0.03	0.81	-6.14	0.84	-4.77	0.57	0.54	1.24
a1_e	1214	436	-0.146	0.05	0.89	-1.72	0.83	-2.30	0.60	0.53	1.23
s3_b	4462	1860	0.268	0.02	0.84	-5.47	0.83	-5.52	0.61	0.57	1.26
d4_e	1081	442	0.220	0.05	0.83	-2.81	0.80	-3.01	0.64	0.56	1.33
d1_c	1366	461	-0.381	0.05	0.85	-2.32	0.80	-2.80	0.58	0.51	1.23
s3_c	4628	1859	0.174	0.02	0.79	-7.28	0.77	-7.33	0.62	0.56	1.33
d2_e	1284	464	-0.125	0.05	0.72	-4.73	0.72	-4.18	0.61	0.54	1.33
d1_e	1186	463	0.097	0.05	0.71	-5.14	0.71	-4.70	0.66	0.55	1.43

*Values larger than limits recommended by Boone et al. (2014)

Table 4.3. Polarity Statistics for Items with Disordered Response Categories

Item Label	Data Code	Data Count	Data Percent[1]	Mean Ability	Mean S.D.	Mean S.E.	Mis- ordered Category	*t*-statistic
a2_g	.	1446	76%	0.66	0.86	0.02		
	1	17	4%	-0.03	0.88	0.22		
	2	13	3%	-0.29	0.53	0.15	*	1.003
	3	68	15%	0.48	0.64	0.08		
	4	96	21%	0.49	0.63	0.06		
	5	258	57%	0.88	0.90	0.06		
e2_a	.	1435	76%	0.65	0.85	0.02		
	1	34	7%	0.21	0.65	0.11		
	2	54	12%	0.20	0.67	0.09	*	0.069
	3	68	15%	0.39	0.68	0.08		
	4	102	22%	0.50	0.72	0.07		
	5	205	44%	1.10	0.92	0.06		
e2_b	.	1430	75%	0.66	0.85	0.02		
	1	14	3%	-0.07	0.68	0.19		
	2	27	6%	0.03	1.05	0.21		
	3	47	10%	0.03	0.36	0.05	*	0.000
	4	123	26%	0.49	0.62	0.06		
	5	257	55%	1.01	0.89	0.06		
e1_d	.	1439	76%	0.67	0.87	0.02		
	1	24	5%	-0.10	0.81	0.17		
	2	19	4%	-0.31	0.57	0.13	*	0.996
	3	78	17%	0.34	0.60	0.07		
	4	128	28%	0.50	0.48	0.04		
	5	210	46%	1.05	0.90	0.06		
e2_c	.	1430	75%	0.66	0.85	0.02		
	1	19	4%	-0.02	0.60	0.14		
	2	32	7%	-0.14	0.74	0.13	*	0.632
	3	81	17%	0.28	0.55	0.06		
	4	129	28%	0.50	0.62	0.05		
	5	207	44%	1.16	0.91	0.06		

[1]*Note.* Missing % includes all categories. Scored % only of scored categories

Table 4.4. Summary of Response Category Structure

Category Label	Observed Frequency	Observed Percent	Observed Average	Sample Expect	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category Measure
1	4675	7%	-0.49	-0.57	1.15	1.44	NONE	(-2.33)
2	8294	13%	-0.05	-0.10	1.07	1.19	-0.90	-0.94
3	13878	21%	0.26	0.31	0.89	0.90	-0.41	-0.05
4	19845	30%	0.73	0.78	0.99	0.89	0.18	0.91
5	19061	29%	1.53	1.48	0.95	0.97	1.13	-2.46

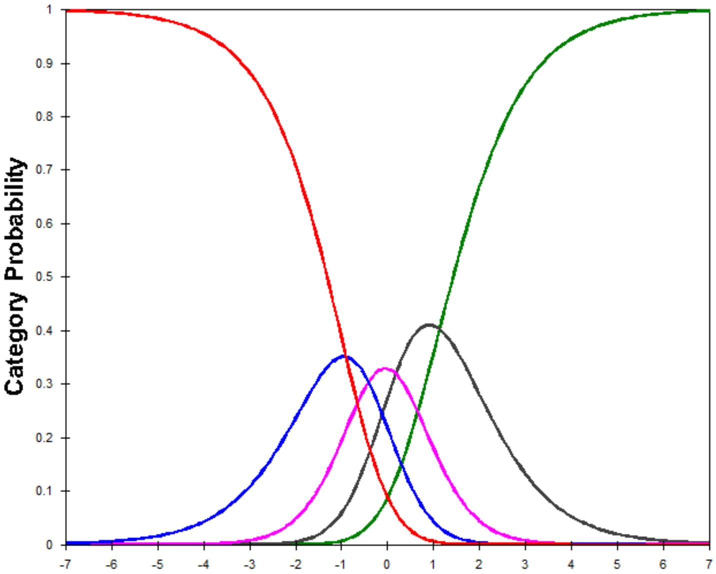


Figure 4.1. Category probability curves for response categories (RSM).
The red line indicates category 1 (Very unlikely), the blue line indicates category 2 (unlikely), the pink line indicates category 3 (neutral), the black line indicates category 4 (likely), and the green line indicates category 5 (very likely).

Theoretical alignment was examined for both scenario difficulty as well as behavior difficulty. Figure 4.2 presents the theoretical-observed alignment of the scenario difficulty. The scenarios are plotted along the y-axis based on hypothetical difficulty, with scenario C2 ranked as the most difficult and scenario S2 ranked as the least difficult. The item difficulties are grouped by scenario and plotted along the x-axis; the mean item difficulty for the scenario is also provided. Scenarios C1 (M = 0.878) and C2 (M = 0.820; i.e., the professor makes a religiously-insensitive joke in class) ranked as the most challenging scenarios, based on average item difficulty. Scenarios E2 (M = -0.536; i.e., a student organization leader posts a racial slur online and the respondent is also a leader of the organization) and S2 (M = -0.486; i.e., the respondent is friends with the woman at the party) ranked as the easiest scenarios, based on average item difficulty. For the most part, the hypothesized scenario difficulty aligns with the observed scenario mean item difficulty. The few exceptions, scenarios B1 (M = -0.287; i.e., possible theft by a known perpetrator) and A2 (M = -0.090; i.e., racial slur in the residence hall when perpetrator and victims are relatively unknown), are considerably misplaced. Table 4.5 provides the mean difficulty along with the number of items, standard deviation, and range of item difficulty within each scenario.

Figure 4.3 presents the theoretical-observed alignment of the behavior difficulty. As with the plot of the scenario difficulty, the possible bystander behaviors are plotted along the y-axis based on hypothetical difficulty, with “Say something at the time to an unknown victim” ranked as the most difficult and “Say something at the time to a known victim” ranked as the least difficult. The item difficulties are grouped by behavior and plotted along the x-axis; the mean item difficulty for the behavior is also provided. Saying something at the time to an unknown

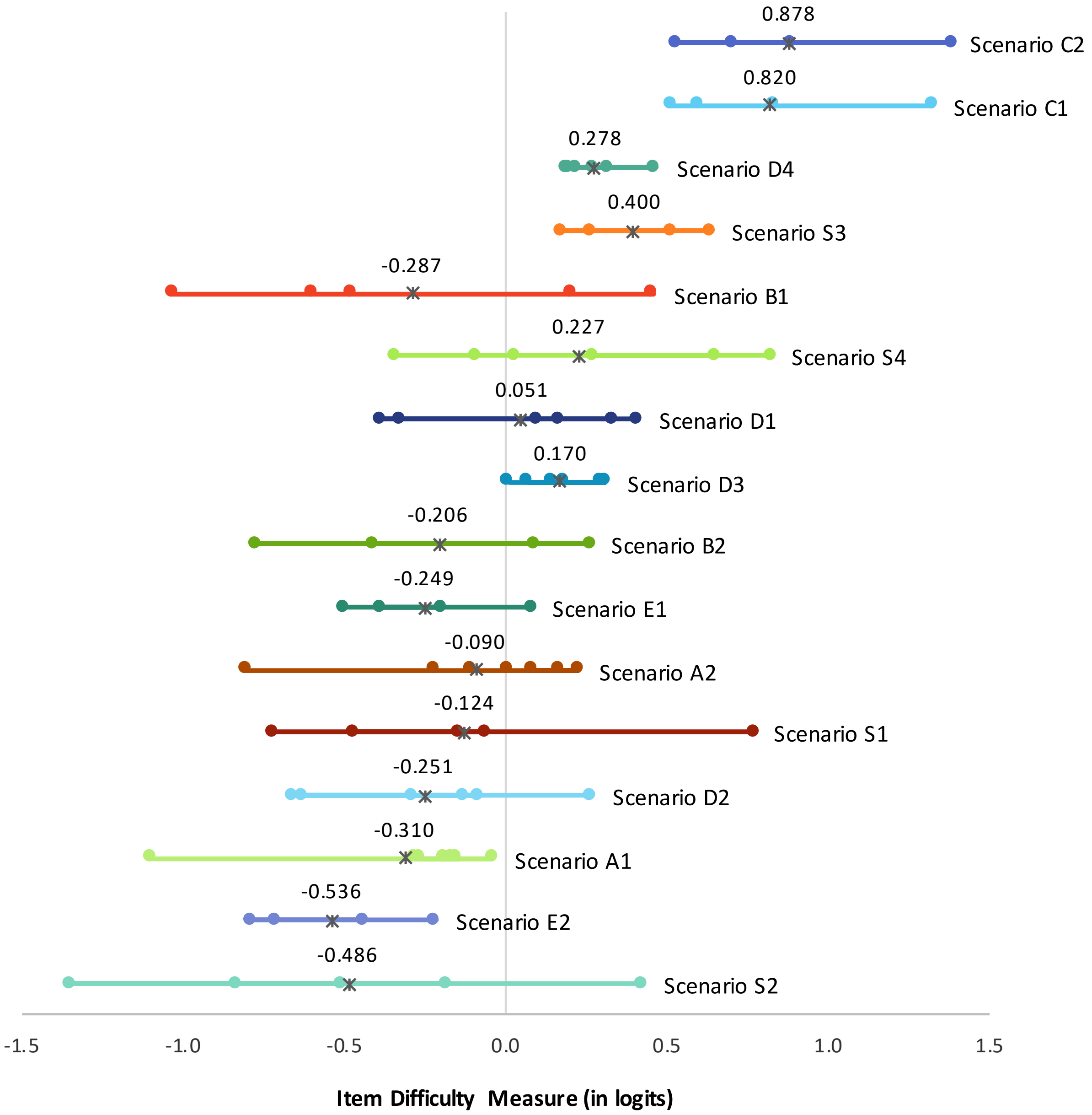


Figure 4.2. Theoretical-observed alignment of average scenario difficulty. Average item ability for each scenario is marked with an “x” and the value is indicated.

victim (*M* = 0.252) and saying something later to an unknown perpetrator (*M* = 0.246) ranked as the most challenging behaviors, based on average item difficulty. Saying something at the time to a known victim (*M* = -0.719) and saying something later to a known victim (*M* = -0.580) ranked as the easiest behaviors, based on average item difficulty. For the most part, the hypothesized scenario difficulty aligns with the observed scenario mean item difficulty. The few exceptions, saying something at the time to a known perpetrator (*M* = -0.497) and saying something later to an unknown victim (*M* = -0.045), are noticeably misplaced. Table 4.6 provides the mean difficulty along with the number of items, standard deviation, and range of item difficulty within each possible behavior.

Table 4.5. Summary Statistics of Scenario Difficulty (Ordered by Observed Mean Difficulty)

Scenario	Hypothesized Difficulty Rank	Number of Items	Mean	SD	Min	Max
Scenario C2	16	4	0.878	0.369	0.531	1.385
Scenario C1	15	4	0.820	0.363	0.516	1.326
Scenario S3	13	4	0.400	0.216	0.174	0.641
Scenario D4	14	6	0.278	0.104	0.192	0.465
Scenario S4	11	6	0.227	0.447	-0.335	0.826
Scenario D3	9	6	0.170	0.119	0.010	0.312
Scenario D1	10	6	0.051	0.333	-0.381	0.410
Scenario A2	6	7	-0.090	0.350	-0.803	0.226
Scenario S1	5	5	-0.124	0.565	-0.718	0.771
Scenario B2	8	4	-0.206	0.472	-0.772	0.264
Scenario E1	7	4	-0.249	0.254	-0.496	0.084
Scenario D2	4	6	-0.251	0.350	-0.653	0.266
Scenario B1	12	5	-0.287	0.610	-1.031	0.459
Scenario A1	3	7	-0.310	0.357	-1.099	-0.036
Scenario S2	1	5	-0.486	0.669	-1.347	0.430
Scenario E2	2	4	-0.536	0.262	-0.786	-0.215

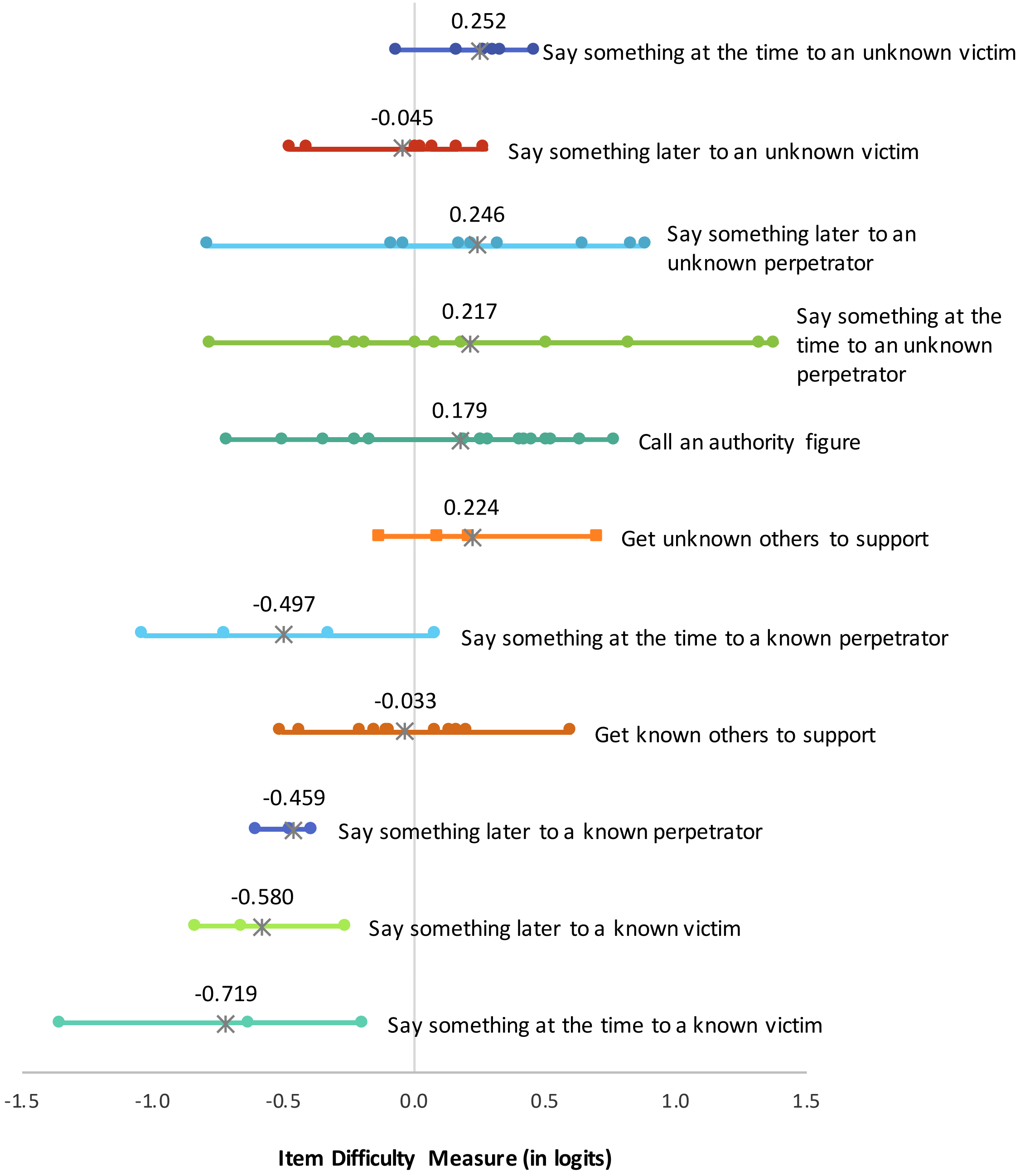


Figure 4.3. Theoretical-observed alignment of average bystander behavior difficulty. Average item ability for each behavior is marked with an “x” and the value is indicated.

Table 4.6. Summary Statistics of Behavior Difficulty (Ordered by Observed Mean Difficulty)

Behaviors	Hypothesized Difficulty Rank	Number of Items	Mean	SD	Min	Max
Say something at the time to an unknown victim	11	7	0.252	0.164	0.465	0.170
Say something later to an unknown perpetrator	9	9	0.246	0.523	0.890	-0.786
Get unknown others to support	6	4	0.224	0.351	0.705	-0.125
Say something at the time to an unknown perpetrator	8	12	0.217	0.670	1.385	-0.772
Call an authority figure	7	16	0.179	0.432	0.771	-0.708
Get known others to support	4	12	-0.033	0.301	0.604	-0.503
Say something later to an unknown victim	10	7	-0.045	0.284	0.273	-0.472
Say something later to a known perpetrator	3	4	-0.459	0.102	-0.381	-0.599
Say something at the time to a known perpetrator	5	4	-0.497	0.484	0.084	-1.031
Say something later to a known victim	2	3	-0.580	0.291	-0.260	-0.653
Say something at the time to a known victim	1	3	-0.719	0.585	-0.188	-0.623

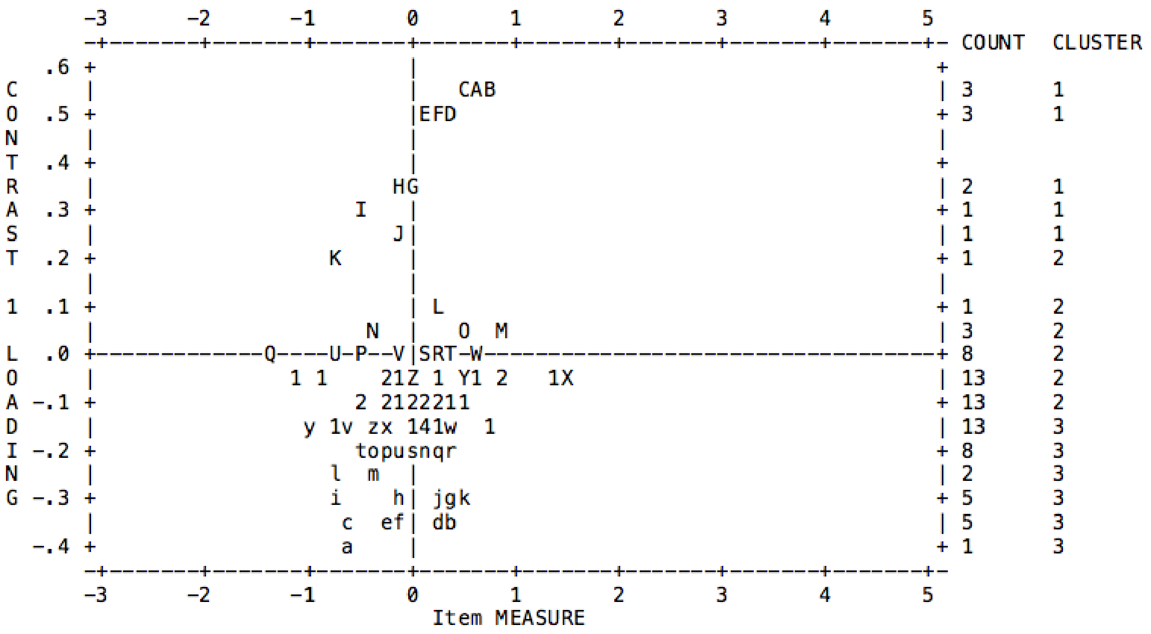
Structural Aspect of Construct Validity

Principle component analysis (PCA) on the standardized residuals from the Rasch scaling of the data was used to test the dimensionality of the instrument. This approach is recommended when the instrument is intended to be unidimensional (Wolfe & Smith, 2007). Table 4.7 presents the eigenvalue units from the PCA on the standardized residuals. The Rasch dimension explains 41.2% of the variance in the data; the items explain 15.0% of the variance. Only two contrasts were provided by Winsteps, with the eigenvalue of the first contrast indicating that there are contrasting patterns in the residuals and the variance in residuals displays structure (Linacre, 2017). This eigenvalue of 4.3 suggests that the contrast has the strength of about four items (out of 83), which is larger than the two items needed for a contrast to be considered a dimension (Linacre, n.d.). This contrast also explains 3.0% of the variance in the data.

Table 4.7. Table of Standardized Residual Variance in Eigenvalue Units

	Eigenvalue	Observed		Expected
Total raw variance in observations	143.3	100.0%		100.0%
Raw variance explained by measures	60.3	42.1%		42.8%
Raw variance explained by persons	38.7	27.0%		27.5%
Raw Variance explained by items	21.5	15.0%		15.3%
Raw unexplained variance (total)	83.0	57.9%	100.0%	57.2%
Unexplained variance in 1st contrast	4.3	3.0%	5.1%	
Unexplained variance in 2nd contrast	3.3	2.3%	3.9%	

Examination of the standardized residual plot suggests a cluster of items are separate from the rest of the items (see Figure 4.4). Table 4.8 presents the item loadings and MNSQ statistics for the six items clustered near the top of the residual plot, as well as the six items at the bottom of the residual plot with which the top cluster of items contrasts. Although the loadings for the items in the cluster are greater than 0.4, the value needed to be considered important by conventional factor analysis, the substance of the items is more important (Linacre, 2012). There does not seem to be a pattern to those items in the cluster, except that they belong to the group of party scenarios. Additionally, the MNSQ values for these items are less than or equal to 1.0, indicating they do not contradict the Rasch model but instead appear to represent a local intensification of the Rasch dimension (Linacre, n.d.).



```
      -3      -2      -1       0       1       2       3       4       5
      -+-------+-------+-------+-------+-------+-------+-------+-------+- COUNT  CLUSTER
   .6 +                        |                                        +
C     |                        |   CAB                                  | 3      1
O  .5 +                        |EFD                                     + 3      1
N     |                        |                                        |
T  .4 +                        |                                        +
R     |                       HG                                        | 2      1
A  .3 +                    I   |                                        + 1      1
S     |                       J|                                        | 1      1
T  .2 +                  K     |                                        + 1      2
      |                        |                                        |
1  .1 +                        | L                                      + 1      2
      |                     N  |   O  M                                 | 3      2
L  .0 +-------------Q----U-P--V|SRT-W-----------------------------------+ 8      2
O     |               1 1    21Z 1 Y1 2   1X                            | 13     2
A -.1 +                    2 2122211                                    + 13     2
D     |                y 1v zx 141w  1                                  | 13     3
I -.2 +                    topusnqr                                     + 8      3
N     |                  l  m  |                                        | 2      3
G -.3 +                  i    h| jgk                                    + 5      3
      |                   c  ef| db                                     | 5      3
  -.4 +                   a    |                                        + 1      3
      -+-------+-------+-------+-------+-------+-------+-------+-------+-
      -3      -2      -1       0       1       2       3       4       5
                                 Item MEASURE
```

Figure 4.4. Standardized residual plot for Contrast 1

Table 4.8. Loading, Measure, and Fit Statistics for Clustered Items

Item Label	Entry Label	Contrast	Loading	Measure	Infit MNSQ	Outfit MNSQ
s3_d	A	1	0.57	0.641	0.93	0.97
s1_e	B	1	0.54	0.771	0.92	1.01
s3_a	C	1	0.54	0.515	0.82	0.84
s2_e	D	1	0.52	0.43	1.01	1.05
s3_c	E	1	0.52	0.174	0.79	0.77
s3_b	F	1	0.51	0.268	0.84	0.83
d2_d	a	2	-0.4	-0.653	1.19	1.08
d4_c	b	2	-0.37	0.323	0.88	0.84
d2_b	c	2	-0.35	-0.623	1.16	1.04
d4_e	d	2	-0.34	0.22	0.83	0.8
d2_a	e	2	-0.33	-0.288	1.03	0.98
d2_c	f	2	-0.33	-0.08	0.95	0.92

Generalizability Aspect of Construct Validity

Three approaches were used to examine the generalizability aspect of construct validity. First, the profile of misfitting persons were examined using chi-square tests of independence to determine if a disproportionate number of misfitting persons belonged to certain demographic groups. Person fit refers to the extent responses perform as the Rash model expects[7] (Boone et al., 2014). Person misfit was evaluated using the outfit ZSTD values, with respondents demonstrating outfit ZSTD greater than 3.0 flagged as underfitting and respondents with outfit ZSTD values less than -3.0 as overfitting; 443 respondents (23.5%) were determined to misfit the Rasch model, with 181 (9.6%) underfitting and 262 (13.9%) overfitting. This result is higher than what would be expected by chance (94 respondents, $\alpha = 0.05$). Table 4.9 presents the results of the chi-square tests. Statistical significance for total misfit was found for worldview/religion ($\chi_4^2 = 18.754, p = 0.001$) and nationality ($\chi_1^2 = 7.840, p = 0.005$). When misfit was disaggregated into underfit and overfit, statistical significance was found for race or ethnicity ($\chi_5^2 = 28.697, p = 0.001$), worldview/religion ($\chi_4^2 = 24.068, p = 0.002$), nationality ($\chi_1^2 = 9.277, p = 0.010$), and academic major ($\chi_5^2 = 19.959, p = 0.030$). Tables 4.10, 4.11, and 4.12 present the frequency and column percentage of misfitting respondents by each race or ethnicity, worldview/religion, and nationality group. These results suggest that Asian/Asian American, Native Hawaiian, or Pacific Islander respondents had higher than expected overfit and Latina/o/x or Hispanic respondents had higher than expected underfit; Black/African American respondents

[7] Since the Rasch model conforms to the requirements of objective measurement, it does not adjust calibrations to the sample used to evaluate an instrument. As such, respondents either fit the model's expectations or not based on the difficulty of the items. When respondents overfit the model, their responses are too predictable. When they underfit the model, their responses are unpredictable. If misfit occurs disproportionately for one group, it may be an indication of an identity-based issue impacting the manner in which the instrument measures the construct.

exhibited lower than expected overfit. Additionally, respondents with minoritized worldviews/religions had higher than expected overfit, whereas worldview majority respondents had higher than expected underfit; nonreligious respondents exhibited lower than expected underfit. Finally, international student respondents had higher than expected overfit, while domestic student respondents had lower than expected overfit.

Table 4.9. Summary of Chi-Square Tests for Misfitting, Underfitting, and Overfitting Persons by Demographic Characteristics

Demographic Group	d.f.	Misfitting Pearson Chi-square	Underfitting and Overfitting Pearson Chi-square
Gender	2	3.132	5.978
Sexual orientation	4	6.582	8.793
Race or ethnicity	5	8.691	28.697***
Worldview/religion	4	18.754***	24.068**
Nationality	1	7.840**	9.277**
Generation status	1	0.798	0.813
Academic class year	5	3.505	6.742
Academic major	5	9.069	19.959*
Institution	2	0.602	0.914
Scenario group	4	2.805	5.391

*p < 0.05, **p < 0.01, ***p < 0.001

Table 4.10. Frequency (Percent) of Fitting and Misfitting Respondents by Race or Ethnicity

	Fitting	Misfitting	Underfitting	Overfitting	Total
Another race or ethnicity	29 (2.0%)	9 (2.0%)	5 (2.8%)	4 (1.5%)	38 (2.02%)
Asian/Asian American, Native Hawaiian, or Pacific Islander	170 (11.8%)	70 (15.8%)	23 (12.7%)	47 (17.9%)	240 (12.7%)
Black/African American	59 (4.1%)	13 (2.9%)	11 (6.1%)	2 (0.8%)	72 (3.8%)
Latino/a/x or Hispanic	88 (6.1%)	36 (8.1%)	20 (11.1%)	16 (6.1%)	124 (6.6%)
Multiracial or Multiethnic	153 (10.6%)	45 (10.2%)	24 (13.3%)	21 (8.0%)	198 (10.5%)
White	943 (65.4%)	270 (61.0%)	98 (54.1%)	172 (65.7%)	1,213 (64.4%)
Total	1,442	443	181	262	1,885

Table 4.11. Frequency (Percent) of Fitting and Misfitting Respondents by Worldview/Religion

	Fitting	Misfitting	Underfitting	Overfitting	Total
Nonreligious	533 (37.0%)	148 (33.4%)	54 (29.8%)	94 (35.9%)	681 (36.1%)
Another worldview	144 (10.0%)	55 (12.4%)	25 (13.8%)	30 (11.5%)	199 (10.6%)
Worldview minority	88 (6.1%)	42 (9.5%)	13 (7.2%)	29 (11.1%)	130 (6.9%)
Worldview majority	598 (41.5%)	190 (42.9%)	85 (47.0%)	105 (40.1%)	788 (41.8%)
More than one worldview/religion	79 (5.5%)	8 (1.8%)	4 (2.2%)	4 (1.5%)	87 (4.6%)
Total	1,735	443	181	262	1,885

Table 4.12. Frequency (Percent) of Fitting and Misfitting Respondents by Nationality

	Fitting	Misfitting	Underfitting	Overfitting	Total
Domestic student	1,394 (96.7%)	415 (93.7%)	172 (95.0%)	243 (92.8%)	1,809 (96.0%)
International student	48 (3.3%)	28 (6.3%)	9 (5.0%)	19 (7.3%)	76 (4.0%)
Total	1,735	443	181	262	1,885

Item difficulty invariance was also used to examine the generalizability aspect of construct validity. Table 4.13 presents the results of the differential item functioning (DIF) analysis. DIF is present for respondent institution, gender, and race or ethnicity. In terms of institution, respondents at the New England university had unusually higher scores on item a1_g than students at the second Far West university (DIF contrast = -0.87, $p < 0.0002$). For gender, items s4_a (DIF contrast = -0.91, $p < 0.0002$) and s4_c (DIF contrast = -0.71, $p < 0.0002$) seem to advantage genderqueer, transgender, or respondents with another gender over cisgender men. Finally, item a2_g seems to advantage Asian/Asian American, Native Hawaiian, or Pacific Islander respondents over white respondents (DIF contrast = 0.74, $p < 0.0001$), whereas item c1_a seems to advantage respondents with more than one race/ethnicity over their Black/African American counterparts (DIF contrast = -0.98, $p = 0.008$).

Finally, person reliability scores were considered as an indicator of internal consistency and evidence of generalizability. Table 4.14 presents the summary statistics of the Rasch analysis, including the real and model person reliabilities. The reliability estimate ranges from 0.92 to 0.94, indicating high internal consistency and instrument reliability.

External Aspect of Construct Validity

Wolfe and Smith (2007) recommended two approaches to examining the external aspect of construct validity: group comparisons and person-item maps. Results from the one-way between-subjects ANOVA indicate group differences in bystander intervention disposition scores for gender (*F*(2, 1439) = 23.24, *p* < 0.001), sexual orientation (*F*(4, 1437) = 4.59, *p* < 0.001), academic major (*F*(5, 1436) = 3.88, *p* = 0.002), and institution (*F*(2, 1237) = 7.11, *p* < 0.001). Given the number of ANOVA tests conducted for this analysis, multiple comparisons and the inflation of Type I error is of concern. As such, only results with significance less than 0.01 are reported. Table 4.15 presents the results from the ANOVAs.

Tables 4.16 through 4.20 present the results from the post-hoc comparisons using Scheffé's test for the demographic characteristics exhibiting differences in the one-way ANOVAs. These results indicate that cisgender women (*M* = 0.86) and genderqueer, transgender, and respondents with another gender (*M* = 0.94) had higher average bystander intervention disposition scores than cisgender men (*M* = 0.54). Furthermore, bisexual respondents (*M* = 0.96) had higher average bystander intervention disposition scores than heterosexual/straight respondents (*M* = 0.69).

For academic major, respondents majoring in the arts and humanities (*M* = 1.07) had higher average bystander intervention disposition than respondents in business administration (*M* = 0.69), the STEM fields (*M* = 0.69), or social sciences or education (*M* = 0.73). Finally, respondents at the second Far West university had (*M* = 0.78) higher average bystander intervention disposition scores than respondents attending the first Far West university (*M* = 0.65) or the university in New England (*M* = 0.57).

Table 4.13. DIF Statistics by Institution, Gender, and Race or Ethnicity for Variant Items

Item Label	Person Class	DIF Measure	DIF SE	Person Class	DIF Measure	DIF SE	DIF Contrast	Rasch-Welch Probability
Institution								
a1_g	Far West 2	-1.77	0.20	New England	-0.90	0.10	-0.87	0.0002
Gender								
s4_a	Cisgender men	0.44	0.04	Genderqueer or another gender	1.36	0.14	-0.91	0.0000
s4_c	Cisgender men	0.38	0.04	Genderqueer or another gender	1.09	0.14	-0.71	0.0000
Race or Ethnicity								
a2_g	Asian/Asian American, Native Hawaiian, or Pacific Islander	-0.23	0.15	White	-0.97	0.08	0.74	0.0000
c1_a	Black/African American	0.61	0.22	More than one race or ethnicity	1.59	0.15	-0.98	0.0008

Table 4.14. Rasch Model Person Summary Statistics

	Raw Score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	126.0	34.9	0.66	0.20	1.06	-0.3	1.07	-0.2
S.D.	26.2	3.7	0.86	0.08	0.66	2.6	0.76	2.4
Maximum	184.0	37.0	5.00	1.11	4.88	8.7	9.90	9.9
Minimum	4.0	1.0	-2.86	0.16	0.00	-6.8	0.00	-6.5
Real Separation	3.31	Real Reliability	0.92					
Model Separation	3.87	Model Reliability	0.94					

Table 4.15. One-Way ANOVA Results for Respondent Demographic Characteristics

Category	Source	SS	*d.f.*	MS	*F*
Gender	Between Groups	36.01	2	18.00	23.24***
	Within Groups	1114.79	1439	0.77	
	Total	1150.80	1441		
Sexual orientation	Between Groups	14.51	4	3.63	4.59***
	Within Groups	1136.28	1437	0.79	
	Total	1150.80	1441		
Race or ethnicity	Between Groups	9.83	5	1.97	2.47
	Within Groups	1140.96	1436	0.79	
	Total	1150.80	1441		
Worldview/religion	Between Groups	3.27	4	0.82	1.02
	Within Groups	1147.52	1437	0.80	
	Total	1150.80	1441		
Academic major	Between Groups	15.34	5	3.07	3.88**
	Within Groups	1135.46	1436	0.79	
	Total	1150.80	1441		
Academic class year	Between Groups	4.90	5	0.98	1.23
	Within Groups	1139.30	1429	0.80	
	Total	1144.20	1434		
Institution	Between Groups	11.27	2	5.63	7.11***
	Within Groups	1139.53	1439	0.79	
	Total	1150.80	1441		
Scenario group	Between Groups	0.83	4	0.21	0.26
	Within Groups	1149.96	1437	0.80	
	Total	1150.80	1441		

**$p < 0.01$, ***$p < 0.001$

Table 4.16. Scheffé's Post-hoc Test of Mean Differences for Bystander Intervention Disposition by Gender

	Mean	Cisgender Man (n = 564)	Cisgender Woman (n = 830)	Genderqueer, Transgender, or Another Gender (n = 48)
Cisgender Man	0.54			
Cisgender Woman	0.86	***		
Genderqueer, Transgender, or Another Gender	0.94	*		

**p* < 0.05, ***p* < 0.01, ****p* < 0.001

Table 4.17. Scheffé's Post-hoc Test of Mean Differences for Bystander Intervention Disposition by Sexual Orientation

	Mean	Bisexual (n = 153)	Gay (n = 32)	Heterosexual/ Straight (n = 1,150)	Lesbian (n = 19)	Queer or another sexual orientation (n = 88)
Bisexual	0.96					
Gay	0.80					
Heterosexual/Straight	0.69	*				
Lesbian	0.66					
Queer or another sexual orientation	0.97					

**p* < 0.05, ***p* < 0.01, ****p* < 0.001

Table 4.18. Scheffé's Post-hoc Test of Mean Differences for Bystander Intervention Disposition by Academic Major

	Mean	Arts and humanities (n = 113)	Business administration (n = 178)	Health Professions (n = 183)	Science, engineering, or mathematics (n = 642)	Social sciences or education (n = 239)	No major selected (n = 87)
Arts and humanities	1.07						
Business administration	0.69	*					
Health Professions	0.80						
Science, engineering, or mathematics	0.69	**					
Social sciences or education	0.73	*					
No major selected	0.67						

**p* < 0.05, ***p* < 0.01, ****p* < 0.001

Table 4.19. Scheffé's Post-hoc Test of Mean Differences for Bystander Intervention Disposition by Institution

	Mean	Far West 1 (n = 537)	Far West 2 (n = 336)	New England (n = 569)
Far West 1	0.73		*	
Far West 2	0.89			
New England	0.66		***	

**p* < 0.05, ***p* < 0.01, ****p* < 0.001

Given the correlation between gender and major selection, a two-way ANOVA with gender and academic major as factors was also conducted to test if the results for academic major was confounded by gender. Table 4.20 gives the results from this test. The interaction effect was not significant ($F(10, 1421) = 1.04$, $p = 0.405$), yet the two main effects were [gender: $F(2, 1424) = 8.41$, $p < 0.001$); academic major: $F(5, 1424) = 3.04$, $p = 0.01$)]. This result suggests that the significant results for differences in academic major are not confounded by gender.

Table 4.20. Two-Way ANOVA Results for Gender and Academic Major

Source	SS	*d.f.*	MS	*F*
Gender	12.93	2	6.47	8.41***
Academic major	11.69	5	2.34	3.04**
Gender x Academic major	8.01	10	0.80	1.04
Within	1094.40	1424	0.77	
Total	1150.80	1441		

***p* < 0.01, ****p* < 0.001

The person-item maps were also used to examine the external aspect of construct validity since these maps indicate discernment of the items. Figure 4.5 presents the Wright map for the respondents and items in this study (the Rasch-Andrich map, which provides the category thresholds is presented in Appendix B). The distribution of respondents along the continuum is relatively normal, with the mean person score higher than the mean item difficulty score.

Additionally, the map shows both floor and ceiling effects, indicating that the items on the instrument do not adequately cover the continuum. Most of the average item difficulty calibrations ranged from -1 to 1 logits; the most difficult item, item c2_a, had difficulty of 1.39 logits, whereas the easiest item, item s2_b, had difficulty of -1.35 logits. These item difficulty values suggest the items exhibited a high degree of construct saturation. In other words, the items differentiated differences in bystander intervention disposition at approximately the same point along the continuum. Table 4.21 presents the item summary statistics. See Appendix C for the item difficulty values as well as the estimated discrimination for each item.

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5             .  +
                  .  |
                  .  |
                     |
                     |
                  .  |
                     |
    4             .  +
                  .  |
                     |
                  .  |
                  .  |
                  .  |
                  .  |
    3             .  +
                  .  |
                  .  |
                 .#  |
                  . T|
                  .  |
                 .#  |
    2            .#  +
                .##  |
                 .#  |
               .### S|
               .###  |  c2
             .#####  |  c1
            .######  |
    1      .#######  +T
        .##########  |  c1  c2  s4
      .############ M|  c2  s1  s4
       .###########  |S c1  c1  c2  s3  s3
      .############  |  b1  d1  d4  s2
         .#########  |  a2  b2  d1  d2  d3  d3  d4  d4  d4  s3  s4
         .#########  |  a2  a2  b1  b2  d1  d1  d3  d3  d3  d4  d4  e1  s3
    0     .########  +M a1  a2  d3  s1  s4
            .###### S|  a1  a1  a1  a2  d2  d2  e1  s1  s2  s4
              .####  |  a1  a1  a2  d1  d2  e2  s4
                .##  |  b1  b2  d1  e1  e1  e2  s1
                 .#  |S b1  d2  s2
                 .#  |  b2  d2  e2  s1
                  .  |  a2  e2  s2
   -1             . T+T b1
                  .  |  a1
                  .  |  s2
                  .  |
                  .  |
                  .  |
                  .  |
   -2             .  +
                     |
                  .  |
                  .  |
                  .  |
                     |
                  .  |
   -3                +
               <less>|<freq>
```

Figure 4.5. Wright map indicating person bystander intervention disposition and item difficulty on the same continuum. Each # represents 7 respondents and each . represents 1-6 respondents.

Table 4.21. Rasch Model Item Summary Statistics

	Raw Score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	2070.2	792.2	0.00	0.05	1.06	0.6	1.10	1.0
S.D.	1639.1	603.4	0.52	0.01	0.19	3.2	0.24	3.5
Maximum	6594.0	1879.0	1.38	0.07	1.76	9.4	1.94	9.9
Minimum	590.0	431.0	-1.35	0.02	0.71	-7.3	0.71	-7.3
Real Separation	10.33	Real Reliability		0.99				
Model Separation	11.01	Model Reliability		0.99				

Conclusion

With Wolfe and Smith's (2007) framework as a guideline to examine five aspects of Messick's (1995) unified concept of construct validity, this study used Rasch modeling to examine the validity of the bystander intervention disposition instrument. Results suggest this instrument demonstrates the content, substantive, structural, and generalizable aspects of construct validity. Additionally, the Rasch analysis indicates this instrument exhibits the external aspect of construct validity, with some limitations. Overall, these results indicate the bystander intervention disposition instrument is valid. In the next section, I discuss these results in further detail before examining the limitations of the study and providing recommendations for future research.

Chapter 5: Discussion

The purpose of this study was to examine an instrument designed to measure this construct specifically within collegiate contexts. Informed by the theories of educational and psychological measurement, this study relied on Rasch analysis techniques to determine the validity and reliability of such an instrument. Analytical methods were organized using Wolfe and Smith’s (2007) framework for mapping Rasch validity evidence to Messick’s (1995) unified concept of construct validity, which includes reliability as an aspect of validity. Should the Rasch evidence conclude this instrument as valid, it can be confidently stated that the instrument measures collegiate bystander intervention disposition and the scores it produces can be interpreted as one’s current level of disposition.

The final chapter of this study fulfills the following purposes: 1) summarize and discuss the findings from the Rasch validity analysis; 2) discuss implications for theory and practice; and 3) consider the opportunities for future research. As such, it is divided into four sections. The first section will interpret the study’s findings, including how the evidence of each construct validity aspect suggests the overall validity of the instrument. This section will also discuss the implications for measuring bystander intervention disposition with the approaches used in this study. The next two sections will discuss the implications for theoretical understanding of bystander intervention and the implications for practice on college and universities. This chapter will conclude with directions for future research.

Summary of Findings

This study utilized Wolfe and Smith's (2007) conceptualization of Rasch validity evidence for Messick's (1995) unified concept of construct validity to organize the methods and results. Five of the six aspects of construct validity were examined, including content, substantive, structural, generalizability, and external. The sixth aspect, consequential, was not considered since this instrument is not intended for use in high-stakes testing in which cut-off scores are required. Taken together, the evidence from the Rasch analysis suggests the collegiate bystander intervention disposition instrument is valid, although areas for improvement exist.

Finding One: Unidimensionality of Bystander Intervention Disposition

The first major finding from this study is although a scenario-based approach was used, these items fit the Rasch model. The Rasch model is a probabilistic model that uses both a person's ability and the difficulty of an item to predict if a person will endorse an item. Item fit refers to the how well items follow the Rasch model's predictions (Boone et al., 2014). When items fit the Rasch model, it not only suggests they represent bystander intervention disposition well, but that they conform to the requirements of objective measurement (Boone et al., 2014; Wright, 1967). Therefore, it can be confidently stated this instrument measures collegiate bystander intervention independent of the sample as well as independent of the items used[8]. Obtaining measurement objectivity allows researchers to "generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements" (Wright &

[8] Sample and item independence indicate the instrument objectively measures the construct, conditions with make it possible to generalize measurement of bystander intervention disposition beyond the specific instrument and sample used to calibrate. Sample independence does not suggest, however, than the instrument will work identically for different populations.

Stone, 1979, p. xii). The fact that these items fit the Rasch model, taken in concert with finding from the dimensionality analysis for the structural aspect of construct validity, additionally suggests that collegiate bystander intervention disposition is a continuous, unidimensional construct which can still be measured across various contexts. In other words, although collegiate bystander intervention disposition is context-related, it is not context-dependent. Other factors, such as cognitive and psychosocial development, may also influence bystander intervention disposition even as the contexts in which bystanders find themselves vary.

Finding Two: Theoretical Alignment

The second major finding relates to the substantive aspect of construct validity: in general, the instrument measures bystander intervention disposition the way it was intended. Results from the item polarity analysis and the category probability curves suggest the rating scale used in this instrument functions the way it was designed: respondents with higher bystander intervention disposition are more likely to say they would engage in bystander intervention behavior when presented with a harmful situation. Respondents are also generally able to discern between the five category responses given for the items. The close proximity of the category thresholds, however, is cause for some concern with the response options. Instead of providing respondents with five options, including a neutral response, it may be better to use a four-category response scale that does not incorporate a neutral option and force respondents to decide whether or not they would engage in a given behavior. Constraining the options may improve the measurement of the items since respondents can no longer remain neutral in a situation nor can they use this response option as a “face-saving way of saying ‘don’t know’” (Sturgis, Roberts, & Smith, 2012, p. 30). Additionally, the lack of statistically significant differences in the mean ability for the five items that demonstrated response category mis-order suggest that respondents struggled to discern between “very unlikely” and “unlikely,” at least for certain behaviors or scenarios. It may also be beneficial to consider in the future a three-point scale with the options “unlikely,” “likely,” and “very likely” for specific scenarios or behaviors.

Examination of the alignment of expected scenario and item difficulty with what was observed also indicates the instrument exhibits the substantive aspect of construct validity. With few exceptions, the observed scenario difficulty – based on the average difficulty of the items in a given scenario – matches what was expected based on past literature. Respondents found the scenarios in which their professor makes a religiously-insensitive joke the most challenging. This result makes sense given the ambiguity of the violence in the situation (i.e., microaggressions contribute to violence directed at minoritized populations, but are often more difficult to notice as violent; see Bollinger & Mata, 2018; Wessler & Moss, 2001), the lack of a clear victim (i.e., the comment was not directed at a specific person), and the perceived authority of the perpetrator. Within these situations, which only vary by the respondent’s relationship with others in the class, the most difficult behavior is saying something in the moment to the professor, followed by saying something to the professor at a later time. It should also be noted that the situation in which the respondent did not have friends in the class was more difficult than if the respondent did have friends present, suggesting that knowing other bystanders in this case made it easier to respond to the violent behavior.

At the other end of the continuum, the scenario in which the respondent was a student leader and needed to address a racial slur made online by another student leader was the easiest in which to intervene, based on average item difficulty. This result aligns with theoretical predictions supported by Banyard, Moynihand, and Crossman’s (2009) and Moynihan and Banyard’s (2008) work. Holding a leadership position within a student organization seems to provide respondents with some authority, and increased sense of responsibility, to stop harmful behavior, even when the perpetrator is another student leader. The influence of holding a leadership role is also reflected by the increased scenario difficulty incurred when the respondent observes the same situation in a position other than student leader. This finding may be explained by the relationship between holding a leadership position and the saliency of one's bystander identity. Student leaders must consider the implications of members' actions for the whole organization, which may increase their willingness to stop harmful behavior. Additionally, as a person within the organization with a level of authority, the costs of intervening may not be as high as they would be for someone who is not in a leadership role. This explanation may clarify why it is more difficult for respondents who are not leaders to intervene.

Additionally, the digital nature of this scenario may also make this situation easier overall (see Brody & Vanelisti, 2016). Since it occurs online and not "in-person," there might be lag in the time when the harmful behavior occurs and when the bystander observes it, giving the bystander more opportunity to recognize the wrongdoing and respond. This point is reflected in the scenario's item-level difficulty. The most difficult behavior related to this scenario is saying something online at the time, whereas the easiest behavior is saying something online later. This finding suggests that when the respondent has time to consider the behavior and formulate a response, they are more likely to address the situation.

The second-easiest scenario for respondents, based on average item difficulty, was the situation in which they are friends with the intoxicated woman seen leaving a party with an unknown man. It was predicted that this scenario would be relatively easy for respondents since they are friends with the potential victim (see Dovidio et al., 2006; Gottleib & Carver 1980; Howard & Crano 1974; Levine et al. 2002), who is clearly entering a dangerous situation with a man the respondent does not know. This supposition of clarity is informed by the many bystander intervention programs that use this exact scenario as an example situation when training collegiate bystanders (see Potter et al., 2009). As such, the easiest item to endorse in this scenario is saying something to the woman at the time, followed by saying something to her at a later time, suggesting that respondents focus on their friendship with the victim in these situations.

The most *surprising* result from this analysis is the ease with which respondents will intervene when someone they know is observed going through a stranger's belongings. It was hypothesized that this scenario would be challenging since bystanders will usually give their friends the benefit of the doubt in these types of situations (see Dovidio et al., 2006). Based on average item difficulty, however, this scenario was the fourth easiest for respondents. Interestingly, the easiest behavior for respondents in this situation is confronting the perpetrator by saying something to them in the moment, followed by saying something to them at a later time. Furthermore, this scenario is easier than a similar one in which the respondent does not know the perpetrator, signaling that the relationship with the perpetrator along with the clarity of the harm done actually encourages respondents to say something.

The possible intervention behaviors were also evaluated for alignment with theoretical predictions. As with the average scenario difficulty, the average intervention behavior difficulty mostly performed as expected. The most challenging intervention behavior was saying something at the time to an unknown victim, whereas the least challenging behavior was saying something at the time to a known victim. This result suggests the relationship with the victim is one of the most important aspects in determining intervention behaviors. It could be that respondents generally felt more responsible for the well-being of their friends (see Brody &

Vangelisti, 2016; Gottlieb & Carver, 1980; Howard & Crano, 1974; Loewenstien & Small, 2007), or the perceived costs associated with helping a stranger (e.g., embarrassment) were high (see Burn, 2009; Latané & Darley, 1970; McMahon & Dick, 2011). The saliency of certain social identities, including a bystander identity, may also shift in ways that support intervention behaviors when respondents encounter people they know (Jones & Abes, 2013). This point is also reflected by the relative ease with which respondents will follow up with unknown victims. Although the term "unknown" is used to describe these possible victims, the fact that the bystander can follow up with them at a later point implies the bystander is familiar with them enough to have the opportunity to say something. It seems even a small amount of acquaintance with the victim will increase the chance the bystander will intervene.

Relationship with the perpetrator also seems to matter to bystanders differently than expected. It was hypothesized that knowing the perpetrator would make saying something more difficult since bystanders could give their friends the benefit of the doubt (see Dovidio et al., 2006) or the associated costs could be high (e.g., making a friend angry, embarrassment, social disapproval; see Dovidio et al., 2006). However, respondents indicated that saying something at the time to a known perpetrator was, on average, the third-easiest intervention behavior. Further, the easiest item in this set was saying something to the known someone going through another person's belongings, followed by saying something to the known man leaving a party with an intoxicated woman. It seems as though bystanders will generally look out for their friends when they do something wrong by trying to stop negative behavior before the situation escalates and gets them into more serious trouble.

Another surprising result from this analysis is the relative difficulty of getting known others to support the bystander intervention behavior. Although it was hypothesized that knowing the other bystanders would encourage getting them to support (Schwartz & Gottlieb, 1980), this finding might actually be a result of the complexity knowing the other bystanders in a situation presents when asking for their help. For instance, if known others do not act, it is more difficult for bystanders to also act, including getting the others to help support intervention (Brody & Vangelisti, 2016; Burn, 2009; Chekroun & Brauer, 2002; Darley & Latané, 1968). On the other hand, knowing the other witnesses to a situation can increase expectations to intervene for everyone (Brody & Vangelisti, 2016; Schwartz & Gottlieb, 1980). Social norms around intervention may be why respondents found getting others to support them easier when the others were known versus when they were unknown (Burn, 2009).

Finally, calling an authority figure, including law enforcement, was a moderately difficult behavior for respondents. Although the costs of involving authorities (i.e., embarrassment for making something a bigger deal than it is, time for follow up later on, etc.), may be perceived as high for respondents generally, it has additional costs for people of color that white bystanders may not consider. Political intersectionality (Crenshaw, 1991) offers some perspective as to why. As Crenshaw (1991) explained,

> Women of color are often reluctant to call the police, a hesitancy likely due to a general unwillingness among people of color to subject their private lives to the scrutiny and control of a police force that is frequently hostile. There is also a more generalized community ethic against public intervention, the product of a desire to create a private world free from the diverse assaults on the public lives of racially subordinated people. (p. 1257)

Linder (2018) additionally noted that women of color who experience sexual violence on campus have to balance their own safety with the safety of their male counterparts who are more likely to experience excessive force by the police and implicit bias by student conduct officers. Relatedly, undocumented students may have a real fear of deportation if they involve law enforcement in potentially harmful situations. So, while university administrators may promote the use of policing as a way to keep campus “safe,” the presence of law enforcement does not necessitate safety for all students. Additionally, students generally do not feel comfortable calling on any law enforcement agency to alleviate a situation, suggesting university administrators should not solely rely on this option for bystander intervention.

In summary, this instrument exhibits the substantive aspect of construct validity; it measures collegiate bystander intervention in the way it was intended. Respondents can discriminate between the five response categories, although a four-category scale will most likely be used in future iterations. Additionally, the observed levels of difficulty demonstrated by the various scenarios and the possible behaviors align generally with what was hypothesized based on theory. However, as this instrument continues to be refined and used in future research, it should be consistently reexamined for this aspect of validity.

Finding Three: Generalizability of Instrument

The third major finding relates to the generalizability of the instrument. Results indicate this instrument exhibits validity across contexts, including respondent characteristics, with some areas for improvement. While 23.5% of the sample misfit the Rasch model, 9.6% underfit and 13.9% overfit the model. This result is higher than what is expected to occur by chance at $\alpha = 0.05$. When respondents misfit the Rasch model, they diverge from what the model expected (Boone et al., 2014). Overfitting the model suggests respondents were too predictable, whereas underfitting the model indicates they were too unpredictable. Linacre (2018) and Boone et al. (2014) offered several reasons for person misfit, including unique personal experiences (e.g., past intervention behaviors), systematic response patterns (e.g., selecting the same response category for a set of items), or rushing through the survey, which can cause unexpected errors at the end. If it seems as though an unexpected number of respondents from certain demographic groups misfit, it might be an indication of an issue in the instrument related to that demographic (Boone et al., 2014).

The results from the chi-square tests of independence indicate that overfitting and underfitting[9] the Rasch model were not independent of race or ethnicity, worldview/religion, or nationality. Upon further inspection, Asian/Asian American, Native Hawaiian, or Pacific Islander respondents overfit the model more than expected, suggesting they disproportionately responded in ways that were too predictable. Latina/o/x or Hispanic respondents, on the other hand, underfit the model more than expected, which signals their responses were disproportionately different from what the Rasch model predicted. Similar findings occurred when the respondent's worldview/religion was examined; worldview minority respondents exhibited higher than expected overfit and nonreligious respondents underfit the model more than expected. International respondents demonstrated higher than expected overfit.

Boone et al. (2014) offered several suggestions as to why someone may misfit the data, including unique personal experiences that influenced the response to an item (or set of items). It may be possible that the saliency of their racial or ethnic identity influenced how these respondents answered the items in ways that differed from their peers with other racial or ethnic social identities. As racial and ethnic minorities on college campuses, these misfitting

[9] Overfitting (too predictable responses) and underfitting (too unpredictable responses) the model does not suggest the instrument was too easy or too hard for certain groups. Additionally, overfitting does not imply the respondents' level of predictability is due to answering the items in a socially desirable way.

respondents may have experienced situations similar to the presented scenarios and have considered how they would respond. In terms of worldview/religion, saliency of one's identity may also be why worldview minority respondents disproportionately overfit the model, while nonreligious respondents underfit. From an academic calendar that breaks for Christmas to the absence of prayer rooms in most buildings to Christian symbols and icons in traditions, most students holding minoritized worldviews (e.g., Islam, Judaism, Hinduism) are constantly aware of the Christian privilege found on many campuses in the United States. Nonreligious students, however, are not subjugated to the same minoritization their worldview minority peers experience. In many cases, nonreligious students simultaneously benefit from Christian privilege while also encountering comfortability in the secular spaces ubiquitous on public campuses. Finally, the overfitting of international respondents may be related to the use of English as the instrument's only language. Boone et al. (2014) offered that reading comprehension may also be one reason for person misfit.

Although these findings may suggest issues with the instrument related to race, worldview, or nationality, the level to which the misfitting occurs is not cause for alarm. Unexpected levels of underfitting, which signals the instrument does not produce predictable scores for a particular group, occurred for only one race or ethnicity as well as only one worldview. It does not seem as though the instrument privileges one group of respondents over another, nor do respondents with majoritized social identities fit the model better than those with underserved or minoritized social identities.

A final point related to person misfit: it is encouraging to see that institution and scenario group did not seem to influence underfit or overfit. These results indicate that the collegiate bystander intervention disposition instrument, as a whole, measures respondents similarly across different institutional contexts. They also signal that the varying scenarios presented to respondents based on their placement in a randomized group also did not influence misfit. Of course, these findings may be limited by the sample used to test the instrument, which is not nationally representative and includes respondents at institutions in historically liberally-leaning locales. Proportion of misfit for both institution and scenario group could change had respondents from more conservative backgrounds been included in the sample.

Item-level invariance was also considered when examining the generalizability aspect of construct validity for this instrument. When an item's difficulty varies across groups, it does not necessarily mean it is unfair to one group over another. Instead these items may define the measurement scale differently for certain groups of respondents. Analysis by institution, gender, and race or ethnicity indicated that five items total exhibited DIF as a function of these three demographic characteristics. Erasing the racial slur directed at a friend in the residence hall was easier, on average, for respondents at the university in New England than it was for the respondents at the second Far West university. The same act when the victim was not a friend of the respondent was easier, on average, for Asian/Asian American, Native Hawaiian, or Pacific Islander respondents when compared to white respondents. Intervening in the domestic violence situation by either saying something to the aggressive student or getting others to support intervening was easier for genderqueer, transgender, or respondents with another gender, on average, than cisgender men. Finally, saying something to the professor in class when friends are present was easier, on average, for respondents with more than one race or ethnicity compared to Black/African American respondents.

Although the presence of DIF does not immediately suggest the instrument is biased, these findings still signal some difference in how students with various social identities respond to the items in this instrument. For instance, Black/African American respondents may not consider correcting a professor in front of the class in the same way as respondents with other racial or ethnic social identities due to the apparent racism still present in college and university life. For these students, the perceived cost of speaking up in these types of situations is just too high. With this in mind, it may make sense to develop culturally targeted instruments which consider the intersectionality of power and privilege for those with various social identities. Moradi and Grzanka (2017), for example, advocate for "critical psychometric work that operationalizes unique manifestations of discrimination shaped by intersections of multiple axes of oppression" (p. 505). As this instrument continues to be tested and used, culturally relevant scenarios and items should be included.

With only five items – out of 83 – demonstrating DIF, it is reasonable to claim the instrument, as a whole, exhibits generalizability validity from this perspective. Further inspection of the items and the groups which exhibit variance across them suggests that the presence of DIF does not imply these items are biased. For the purposes of this study, these items were not removed from the instrument, although the presence of DIF for these items is a limitation for the subsequent analysis. As the collegiate bystander intervention disposition instrument is refined, examination of DIF should continue to ensure quality of the items across multiple groups and objective measurement using the instrument.

Finally, person reliability values suggest strong internal consistency for the items on the collegiate bystander intervention disposition instrument. The reliability estimate, which can be interpreted similarly to Cronbach's alpha, ranges from 0.92 to 0.94, implying high instrument reliability. This result also supports the assertion that this instrument exhibits the generalizability aspect of construct validity.

In closing, results from the Rasch analysis provide enough evidence to suggest this instrument exhibits the generalizability aspect of construct validity. When tested for person misfit, a disproportionate number of respondents who identified as Asian/Asian American, Native Hawaiian, or Pacific Islander overfit the model, whereas a disproportional number of Latina/o/x or Hispanic respondents underfit the model. Similar results for worldview/religion and nationality were observed. However, these findings do not indicate the instrument privileges one group of respondents, namely those with majoritized social identities, over others. Additionally, the presence of DIF for only a few items suggests the items on the survey are generalizable across multiple contexts and social identities. However, the generalizability aspect of construct validity should be reassessed as the instrument is used in future studies and as it is refined for cultural relevancy.

Finding Four: Levels of Bystander Intervention Disposition

The fourth and final major finding is the level to which the collegiate bystander intervention disposition instrument exhibits the external aspect of construct validity. One-way between-subjects analysis of variance indicated significant differences in bystander intervention disposition scores for respondents' gender, sexual orientation, race or ethnicity, academic major, and institution. Post-hoc comparisons reveal that cisgender women and genderqueer, transgender, or respondents with another gender had higher bystander intervention disposition, on average, than cisgender men. This finding aligns with what was predicted (see Bennett, Banyard, & Edwards, 2015; Brown, Banyard, & Moynihan, 2014; Burn, 2009; Tjaden & Thoennes, 2000) and provides evidence for the external aspect of validity.

Relatedly, bisexual respondents exhibited more bystander intervention disposition, on average, than heterosexual/straight respondents. Since little research examining the role sexual orientation plays in bystander intervention behaviors currently exists, it is difficult to explain this finding. Although on the surface it may seem that bisexual respondents have higher intervention disposition due to the marginalization of their sexual identity, it is most likely more complex than that. As individuals whose social identity does not fit neatly into a single category, they experience violence in an intersectional way (Crenshaw, 1991). For instance, Paul (1984) referred to bisexual identity as "an idea without social recognition" (p. 45), noting that for bisexual individuals, "as they are not fully integrated into any one group, there is no group from which they are not to some extent deviant" (p. 53). In other words, many bisexual individuals experience double marginalization as they attempt to belong to both the heterosexual and gay and lesbian communities. Furthermore, researchers have documented worse mental health outcomes, including increased incidents of anxiety, depression, self-harming, and drug use in bisexuals (Flanders, Dobinson, & Logie, 2017; Jorm, Korten, Rodgers, Jacomb, & Christensen, 2002; Li, Dobinson, Scheim, & Ross, 2013), with bisexual women most susceptible to sexual assault in college (Ford & Soto-Marquez, 2016). It could be that bisexual respondents understand these challenges facing their community, and perhaps may be more likely to intervene in situations where they see others at risk for victimization. However, more research is needed to fully explore and understand this interesting finding.

One of the more surprising findings from this analysis is the lack of significant differences in average bystander intervention disposition for race or ethnicity. Although the one-way ANOVA indicated trending differences in average disposition among the race or ethnicity categories, the results are currently inconclusive. This finding differs from what was predicted (see Brown, Banyard, & Moynihan, 2014; Kunstman & Plant, 2008). One limitation to this analysis in particular is the disproportionate number of Asian/Asian American, Native Hawaiian, or Pacific Islander respondents and Latina/o/x or Hispanic respondents misfitting the Rasch model. The ANOVA was conducted using only data which fit the model, so those misfitting respondents were not included in this analysis. Another reason for this unexpected result may be the cultural relevance of the scenarios for respondents of various racial and ethnic backgrounds. As with the misfit and DIF analysis, the development of culturally targeted measures should be considered if further research is conducted on this topic.

The fact that respondents majoring in the arts and humanities had higher bystander intervention disposition, on average, than respondents in business administration, the STEM fields, and the social science or education – even after controlling for gender – also aligns with theoretical predictions (see Abbott & Cameron, 2014; Banyard et al., 2016; Brinkman et al., 2015). Not only do students in arts and humanities majors focus on critical thinking and integrative learning, they may also engage in discussions about diversity and opportunities for reflection more often than students in these other majors which could lead to increased bystander intervention disposition (see Mayhew & DeLuca Fernández, 2007). It is surprising, however, that class year was not a factor related to bystander intervention disposition. Given the increased self-authorship that occurs over the four years of attending college and the strengthened sense of belonging associated with more time on campus (see Banyard et al., 2016; Carlo & Randall, 2001), it was expected that class year would make a difference in bystander intervention disposition. Further studies should, however, continue to look at this relationship.

Finally, the post-hoc test for institution revealed that respondents attending the second Far West institution had higher bystander intervention, on average, than respondents at the first Far West university as well as respondents at the university in New England. Further investigation of the institutional practices and past campus incidents might account for why this is the case. Additionally, there was not a significant difference in the bystander intervention disposition scores for respondents in the four randomly assigned scenario groups. This result is positive and gives further evidence supporting the fact that the scenarios a respondent was given did not influence their overall bystander intervention disposition. However, as with the misfit results, this finding is limited by the sample used in the study.

The final piece of evidence for external validity is the person-item map, which places the person scores along the same continuum as the item difficulty. This map provides information regarding the distribution of item difficulty along the continuum as well as shows where respondents are clustered based on their measurement scores. For analysis using the RSM, two maps are typically examined: The Wright map and the Rasch-Andrich Map. The Wright map, which plots item difficulty with respondent disposition, showed most of the items on the collegiate bystander intervention disposition instrument were relatively easy for respondents; the mean item difficulty was nearly 1 logit lower than the mean person disposition. The map also indicated that most of the items were redundant and measured at the same difficulty level. Floor and ceiling effects were also present, suggesting that easier and harder items are needed. Further inspection of the Rasch-Andrich map, which maps the difficulty of each item response category instead of the overall item difficulty, suggests better coverage of the continuum by the items. It also shows the redundancy of the items. In summary, many of the current scenarios and items measure bystander intervention disposition similarly and can be removed, while new more difficulty scenarios and items are needed to span the continuum.

There are a few possible reasons why the items and scenarios all measure at approximately the same level. The first is the generality of the scenarios. The scenarios were written to be demographically neutral; other than the party scenarios in which the intoxicated student is described as a woman and the sober students is described as a man, none of the other scenarios provide the gender, sexual orientation, race or ethnicity, or worldview characteristics of those in the situation. The decision was made to make the situations free of these demographics to avoid stereotyping certain groups. However, this neutrality means that respondents can "imagine" whomever they want, which means they could possibly think of the actors as people who look like them or as the stereotypical person in that role. For instance, nothing is known about the professor's social identities in the scenario with the religious microaggression in class. Would it matter to the respondents if this professor is a man or a woman, if they are white or a person of color, or if they hold a minoritized religious identity? And what about the other student bystanders in the class? Do their identities matter as well? Finally, would the religion the insensitive joke references change how respondents view the situation? Intersectionality theory (Crenshaw 1989, 1991) would suggest that the way others perceive violence is different based on their social identities and the overlapping systems of power that differentially impact people based on their social location. It is possible that adding these types of identity details could change the scenario's level of difficulty. However, this change in difficulty could happen for some respondents and not others; should this change be made in the future, DIF should be regularly examined and considered.

Another reason these items seem to measure collegiate bystander intervention disposition at the same level is that relationships to others in a scenario does not make as much of a difference in the level of difficulty as other factors. For instance, the set of four D-scenarios (i.e., the streaming of a roommate's sexual encounter) had the most variation in terms of relationship to the perpetrator, victim, and other bystanders, yet, the three scenarios in which the respondent does not know the victim in the video have similar average difficulty (see Table 4.5). This result, along with the others, suggests the one relationship that seems to matter the most is relationship with the victim. Nearly all the scenarios in which the respondent knows the victim are the easiest to endorse, based on average item difficulty. It may be more beneficial for future iterations of this survey to vary the relationships with the victim instead of including different relationships with the perpetrator or other bystanders.

Additionally, although the continuum of violence (Kelly, 1987, 1989) was reflected in the instrument, it could have been more explicit in the scenarios. For instance, none of the scenarios include physical violence other than the domestic altercation in which physical violence is implied. Furthermore, the continuum of violence is not explicit for any particular identity group. The microaggression in class is religious in nature, whereas the slur directed at a peer in the residence hall is racial. It could possibly make the situations more difficult if the continuum of violence was reflected for each minoritized group.

Although additional items are needed to increase the difficulty of the instrument, these maps do suggest the collegiate bystander intervention disposition instrument functions well, generally speaking. Most of the respondent scores fall between two standard deviations of the mean, suggesting a relatively normal distribution. Further, there are no major gaps in the instrument's ability to measure along the continuum when the response categories are considered, and the item separation is high. These results, along with the person separation values found in Table 4.14, indicate the instrument is able to distinguish respondents' level of disposition and the items can be differentiated from each other.

Summary

This instrument is a valid first attempt to measure this newly defined construct of collegiate bystander intervention disposition. The Rasch evidence supports the content, substantive, structural, generalizable, and external aspects of Messick's (1995) unified concept of construct validity. The items on this instrument measure a single, continuous latent construct, even though they were scenario-based. The observed scenario and bystander behavior difficulties, based on average item difficulty, align with what was hypothesized, and the rating scale functions the way it was intended. DIF analysis revealed only a few variant items and the person reliability values were high. Significant differences in disposition scores were also matched what was hypothesized for demographic groups and the Rasch-Andrich map shows acceptable coverage of the continuum, once category response thresholds are considered.

The short answer to the research question, "Is this instrument a valid way to measure collegiate bystander intervention?" is yes. But this instrument is not without its shortcomings and areas for improvement. In the next section, I discuss the contributions to theory, the implications for practice, and the recommendations for future research.

Contributions to Theory

The conceptualization of bystander intervention disposition outlined in this study contributes to a new way of thinking about collegiate bystanders, the situations they observe, and how they come to their intervention behaviors to the theories of bystander intervention. Although theories of violence, bystander cognition, and bystander psychosocial development were used to demonstrate the validity of this instrument, the findings from this study also informs the future directions for theoretical understandings of collegiate bystander behaviors. This section outlines these contributions by discussing how this construct and instrument changes the perception of bystander decision-making.

First, most of the bystander theories and literature have focused on the cognitive processes by which bystanders make their decisions separately from the environmental context in which bystanders find themselves. Latané and Darley's (1970) Decision Model of Helping remains the most common theory used to frame how bystanders decide what actions to take in situations. Yet, bystander intervention disposition is comprised of other domains as well, namely the environmental context and identity development. As this measure demonstrates, intervention decisions are not just cognitive in nature, but are made in concert with perceptions of the environment and saliency of social identities. Given the importance of social identities in determining one's perception of violence and the intersectionality of violence for those holding multiple marginalized identities (see Crenshaw, 1991), environmental context and identity saliency should be considered in future theories of bystander intervention.

This study also contributes to an understanding of what situations and actions collegiate bystanders find easy and which ones they find challenging. The situations and behaviors in which the respondents knew the possible victim were the easiest to endorse. Additionally, it was relatively easy to engage in bystander behaviors when respondents knew the possible perpetrator. These findings add some nuance to the cost-reward model developed by Piliavin et al. (1981) and Dovidio et al. (2006). Instead of the rational evaluation of costs and benefits to the bystander and the victim, bystanders seem to consider their relationship with the victim and perpetrator the most when it comes to deciding to intervene. It may be that bystanders feel an increased sense of responsibility for their friends that overrides the costs of helping, which suggests a morality-influenced decision (see Hoffman, 2000). These respondents also may feel more comfortable interacting with their peers than with outsiders, even if the peer is the perpetrator, suggesting that how potential costs are evaluated by the bystander may not be as clear cut as originally posited (i.e., embarrassment, lost friendship, etc.). Finally, this finding may accentuate the importance of social norms in bystander behaviors. It may be that bystanders are more likely to call out or look after their friends because they have established social norms within their peer networks that support such behaviors.

Finally, theorists tend to consider bystanders monolithically. They should not. In the same way researchers need to produce culturally-relevant instruments for bystanders with multiple marginalized identities, theorists should also begin to include perspectives of social identities and intersectionality into their frameworks. This study demonstrates that the ways one perceives their social identities, and the overlapping systems of power and oppression based on the social location of these identities (i.e., intersectionality; Crenshaw, 1991), are important factors to consider when discussing bystanders and bystander intervention. For instance, there seems to be an increasing trend of whites calling law enforcement on Blacks and other people of color for seemingly mundane behaviors that would not result in police intervention had the "perpetrators" been white (e.g., using a neighborhood pool, barbequing in a public park, watching a child's soccer game, sitting in Starbucks, entering their apartment building, etc.; see O'Donnell, 2018; Molina, 2018; Noori Farzan, 2018). Although these white bystanders may tell themselves they are intervening in some sort of wrong-doing to justify their actions, they are actually contributing to racial violence and the systems of oppression that continue to impact people of color in the United States. As Johnson (2018) pointed out,

> Calling the police is the epitome of escalation, and calling the police on black people for noncrimes is a step away from asking for a tax-funded beatdown, if not an execution… The intent of these actions is to remind black people that the ultimate consequence of discomforting white people—let alone angering them—*could be death*. (para. 14)

So, the question must be asked: At what point do these white bystanders actually become perpetrators of racial violence? As scholars continue to examine harmful situations and the bystanders who observe them, they need to interrogate the intent from the impact of such behaviors.

Implications for Practice

The results from this study shed some light on the policies and programs college and university administrators can implement to create an environment that supports the development of bystander intervention disposition. First, since students are already trained to notice and stop campus sexual violence, they should also participate in educational programs that equip them with the knowledge and skills to detect and respond to other forms violence on campus, including identity-based violence. These initiatives should include an intersectional perspective and instruct collegiate bystanders with privileged social identities to examine their own biases and assumptions while remaining culturally relevant for all students. Additionally, robust instruments – like the one developed in this study – should be utilized to evaluate the efficacy of such programs. Finally, faculty and staff should also be trained to both observe and react to these types of situations as well as support student bystanders who report instances of violence to them.

As a supplement to increased training, college and university administrators should also implement programs with the intent of building community on their campuses. Relationships with the victim and the perpetrator where both factors that made intervention easier for collegiate bystanders, so providing students with the opportunities to expand their peer networks may increase bystander intervention disposition. Since these types of programs and activities occur naturally in extra- and co-curricular spaces (e.g., residence life, student activities, collegiate recreation), faculty should be encouraged to build community among students within their classroom spaces. The ways faculty structure their teaching and pedagogy are also related to bystander intervention disposition. Respondents in arts and humanities courses had higher bystander intervention disposition than students in business administration, the STEM fields, and social sciences, suggesting that approaches traditionally found in these disciplines – discussions about socio-cultural issues, opportunities to integrate learning across subjects, and emphasis on thinking critically – may influence bystander intervention disposition if used in other fields of study.

Finally, college and university administrators should consider implementing those policies which encourage all types of bystander intervention behaviors. For instance, campuses could train law enforcement and student conduct officers in intersectionality theory in ways that restore trust in their ability to keep all members of campus, including those with multiple marginalized identities, safe. Another possible direction is the implementation of a group of responders separate from traditional law enforcement that can support collegiate bystanders who may not feel confident in their ability to stop violence through direct intervention, but also are not comfortable relying on current resources. Campus administrators may also want to develop policies that protect students who step in to seek medical help for peers who may be dangerously intoxicated. Similar policies should also be established for students who report instances of bias by faculty and staff. No collegiate bystander should feel the costs of intervention are too high because of campus policies or procedures.

Recommendations for Future Research

Future direction for this research falls into one of three categories: current instrument refinement, other constructs related to bystander intervention that could be measured similarly, and exploration of collegiate environments using a refined instrument. These three future lines of inquiry are discussed in this section.

As previously noted, there are several aspects of the current instrument that should be changed for the next iteration. Several of the current scenarios and items are redundant, while harder scenarios and items are needed. The scenarios used on the next version should also include more details regarding the perpetrators, victims, and other bystanders. These details could include information on their social identities, which may intersect in various ways. However, adding this information should be done with care as to not revert to stereotypes of marginalized populations. Additionally, the continuum of violence should be reflected more explicitly in the situations provided. I also think it would be beneficial to use a four-point response scale which ranges from "very unlikely" to "very likely" with no neutral point.

One interesting direction I have considered with this instrument is focusing on bystander intervention disposition as it relates to identity-based violence. Hate crimes motivated by race, religion, sexual orientation, gender, ethnicity, national origin, and disability on college and university campuses have increased substantially in the last two years (US DOE, 2018). Although I agree that sexual and partner violence should continue to be measured and addressed, I believe identity-based violence should be considered as well. Narrowing the scope of this instrument may also improve its ability to measure along the continuum since many of the scenarios in the current instrument are identity-based. Additionally, the lack of significant differences in bystander intervention disposition by race or ethnicity may suggest that future iterations of this instrument should include culturally relevant and targeted scenarios and items. Given the Rasch model's capability of handling missing data, a similar approach to the one used in this study could be used by which respondents of various social identities receive comparable, but more applicable situations and questions. Similarly, I think it might be useful to focus on in-person scenarios separate from on-line ones. This refined instrument would now measure collegiate bystander intervention disposition for in-person, identity-based violence. A draft version of the possible next instrument is found in Appendix D.

I also would like to continue to explore the validity of instruments that measure bystander intervention disposition in situations of sexual and partner violence. As the review of the current instruments that measure sexual and partner violence bystander intervention attitudes, self-efficacy, and myths in this study suggests, there is a need for a psychometrically valid instrument that measures bystander intervention disposition for these types of situations. Relatedly, it could also be interesting to use this scenario-based approach to measure bystander intervention-like constructs. This study conceptualized bystander intervention within the context of violence, but there are other conceptually different situations in which people may need to intervene. For instance, witnessing unethical behaviors within an organization, responding to accidental harm, and other prosocial behaviors. Further research could consider the method used in this study when measuring these similar constructs.

Finally, this instrument was designed and tested in hopes it would be used to examine those collegiate environments which increase, or decrease, collegiate bystander intervention disposition. College-going has been associated with a number of outcomes related to bystander intervention (see Mayhew et al., 2016). Since the start of the 21st century, students have demonstrated substantial increases in leadership, self-concept, and independence from authority during college, although more information is needed to understand if these changes differ for those who have not attended college. The overall collegiate environment also influences civic tendencies, with religiously affiliated colleges and universities reinforcing civic outcomes and students' aggregate religious attributes predicting civic behavior.

Within colleges, diversity-focused coursework has been associated with promoting "awareness of other ethnic groups and cultures, openness to diversity, desire to promote racial understanding, and dispositions toward community and civic engagement" (Mayhew et al., 2016, p. 550). Additionally, service-learning courses appear to increase students' commitments to social justice and advance their sense of social responsibility and civic engagement. These findings, however, do not appear to be conditional on student characteristics such as race, gender, worldview, and sexual orientation. Attending college is also associated with increased community and political engagement after college.

Despite these findings, little is known about how college-going actually affects prosocial behaviors generally and bystander intervention specifically. Emerging evidence supports the efficacy of bystander intervention programs to equip college students with the knowledge, awareness, and skills needed to intervene in situations of sexual violence (see Katz & Moore, 2013), but no other claims can be made about these programs. Do they actually change bystander intervention disposition across multiple contexts? Do they elicit change in bystander behaviors in the long-term? Now that an instrument has been developed to measure bystander intervention disposition, more research can be conducted to understand the influence of college-going on bystander intervention disposition and prosocial behaviors more broadly.

Conclusion

The purpose of this study was to evaluate the validity of an instrument designed to measure collegiate bystander intervention disposition. Disposition was selected to describe this psychological construct as it reflects one's character, state of readiness, and/or tendency to act in a specified way that can also be learned (Bourdieu, 1990). Bystander intervention disposition is therefore considered a latent construct that influences behavior and can be changed through interaction with educational settings.

This study employed several theoretical perspectives to conceptualize bystander intervention disposition. As an acquired system, disposition is influenced by one's context as well as one's personal characteristics (Bourdieu, 1990, Bronfenbrenner, 1979, 1993). Therefore, the study of bystander intervention disposition was informed by current understandings of violence, cognition and moral reasoning, and identity. The socio-ecological theory of violence (Dahlberg & Krug, 2002), the continuum of violence (Kelly, 1987, 1989), and intersectionality (Crenshaw, 1989, 1991; Moradi & Grzanka, 2017) each contributed to a holistic understanding of the types of situations bystanders should attempt to interrupt. Theories of bystander decision-making processes, such as Carlo and Randall's (2001) Socioecological Developmental Model of Prosocial Action, Latané and Darley's (1970) Decision Model of Helping, Piliavin and colleagues' (1981) cost-reward perspective, and Hoffman's (2000) theoretical framework of moral reasoning centered the development of empathic distress, informed the cognitive aspect of bystander intervention disposition. Finally, the RMMDI (Jones & Abes, 2013), which draws on theories of meaning-making (Kegan, 1994) and self-authorship (Baxter Magolda, 2001, 2009) to reconceptualize identity development, was instrumental in considering how social identities influence bystander intervention disposition as well as how meaning-making and self-authorship influences one's bystander identity.

The instrument used in this study resulted from of the qualitative work on campus violence conducted by Mayhew, Caldwell, and Goldman (2011) and the psychometric testing of similar items by Mayhew, Lo, Dahl, and Selznick (2018). The instrument contained 16 scenarios which differed by context, violent behavior, and relationships with the perpetrator, victim, and other bystanders. For each scenario, a series of bystander intervention behaviors were presented with respondents asked to rate their likelihood of engaging in that behavior on a five-point Likert-type scale (1 = “Very unlikely” to 5 = “Very likely”). Across the 16 scenarios, there were 83 items total.

The validity and reliability of this instrument was tested on a sample of 1,939 students at one of three universities in the United States. Respondents all saw the same four scenarios at first, but then were randomly assigned to one of four groups to see three additional scenarios; these additional scenarios were different for each of the four groups. Wolfe and Smith’s (2007) Rasch validity framework was adopted to examine the psychometric functions of the instrument, with various Rasch analytical techniques used to evaluate five aspects of Messick’s (1995) unified concept of construct validity. Results from the Rasch analysis suggested this instrument was a valid and reliable way to measure collegiate bystander intervention disposition.

As higher education administrators continue to fight campus violence in its many forms, this study provides important insights into the measurement of bystander intervention disposition broadly. Now that bystander intervention disposition can be validly and reliably measured, educators and researchers can begin to assess how college and university students develop along this continuum and the institutional educational experiences that contribute to this development. The final hope of this study is to help make higher education institutions more equitable and pluralistic for all community members.

References

Abbey, A., Ross, L. T., McDuffie, D., & McAuslan, P. (1996). Alcohol and dating risk factors for sexual assault among college women. *Psychology of Women Quarterly*, *20*(1), 147-169.

Abadu, M. (1991). Current trends in diversity affecting American higher education. *Black Issues in Higher Education, 8*(13), 46-55.

Abbott, N., & Cameron, L. (2014). What makes a young assertive bystander? The effect of intergroup contact, empathy, cultural openness, and in-group bias on assertive bystander intervention intentions. *Journal of Social Issues*, *70*(1), 167–182.

Abes, E. S., Jones, S. R., & McEwen, M. K. (2007). Reconceptualizing the model of multiple dimensions of identity: The role of meaning-making capacity in the construction of multiple identities. *Journal of College Student Development*, *48*(1), 1-22.

Abumrad, J., & Krulwich, R. (2018, Janurary 9). *How to be a hero*. RadioLab Podcast. Podcast retrieved from http://www.radiolab.org/story/how-be-hero/

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, *32*(4), 665-683.

Amar, A. F., Sutherland, M., & Kesler, E. (2012). Evaluation of a bystander education program. *Issues in Mental Health Nursing*, *33*(12), 851–857.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Armor, D.J. Theta reliability and factor scaling. In H.L. Costner (Ed.), *Sociological methodology* (pp. 17-50). San Francisco: Jossey-Bass.

Aultman, B. (2014). Cisgender. *Transgender Studies Quarterly, 1*(1-2), 61-62.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bannon, R. S., Brosi, M. W., & Foubert, J. D. (2013). Sorority women's and fraternity men's rape myth acceptance and bystander intervention attitudes. *Journal of Student Affairs Research and Practice, 50*(1), 72-87.

Banyard, V. L. (2008). Measurement and correlates of prosocial bystander behavior: The case of interpersonal violence. *Violence and Victims, 23*(1), 83-97.

Banyard, V. L. (2011). Who will help prevent sexual violence: Creating an ecological model of bystander intervention. *Psychology of Violence, 1*(3), 216–229.

Banyard, V. L. (2015). *Toward the next generation of bystander prevention of sexual and relationship violence*. Cham, Switzerland: Springer International Publishing.

Banyard, V. L., & Moynihan, M. M. (2011). Variation in bystander behavior related to sexual and intimate partner violence prevention: Correlates in a sample of college students. *Psychology of Violence, 1*(4), 287–301.

Banyard, V. L., Moynihan, M. M., & Plante, E. G. (2007). Sexual violence prevention through bystander education: An experimental evaluation. *Journal of Community Psychology, 35*(4), 463–481.

Banyard, V. L., Moynihan, M. M., Cares, A. C., & Warner, R. (2014). How do we know if it works? Measuring outcomes in bystander-focused abuse prevention on campuses. *Psychology of Violence, 4(*1), 101–115.

Banyard, V. L., Plante, E. G., & Moynihan, M. M. (2004). Bystander education: Bringing a broader community perspective to sexual violence prevention. *Journal of Community Psychology, 32*(1), 61–79.

Banyard, V. L., Plante, E. G., & Moynihan, M. M. (2005). *Rape prevention through bystander education: Bringing a broader community perspective to sexual violence prevention*. Final report to NIJ for grant 2002-WG-BX-0009. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/208701.pdf

Banyard, V., Weber, M. C., Grych, J., & Hamby, S. (2016). Where are the helpful bystanders? Ecological niche and victims' perceptions of bystander intervention. *Journal of Community Psychology, 44*(2), 214–231.

Barnett, N.D., & DiSabato, M. (2000). Training camp: Lessons in masculinity. In J. Gold & S. Villari (Eds.), *Just sex: Students rewrite the rules on sex, violence, activism and equality* (pp. 197-210). Lanham, MD: Rowman & Littlefield.

Basile, K. C., & Smith, S. G. (2011). Sexual violence victimization of women: Prevalence, characteristics, and the role of public health and prevention. *American Journal of Lifestyle Medicine, 5*(5), 407-417.

Batson, C. D. (1998). Altruism and Prosocial Behavior. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 282–316). New York, NY: McGraw-Hill.

Baum, K., & Klaus, P. (2005). Violent victimization of college students, 1995-2002. *Bureau of Justice Statistics Special Report.* Washington, DC: U.S. Department of Justice. Retrieved from https://www.bjs.gov/content/pub/pdf/vvcs02.pdf

Baxter Magolda, M. B. (2001). A constructivist revision of the measure of epistemological reflection. *Journal of College Student Development*, *42*(6), 520-34.

Baxter Magolda, M. B. (2009). The activity of meaning making: A holistic perspective on college student development. *Journal of College Student Development*, *50*(6), 621-639.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300.

Bennett, S., Banyard, V. L., & Edwards, K. M. (2017). The impact of the bystander's relationship with the victim and the perpetrator on intent to help in situations involving sexual violence. *Journal of Interpersonal Violence*, *32*(5), 682–702.

Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology*, *12*(5), 409-415.

Berg-Cross, L., Starr, B. J., & Sloan, L. R. (1994). Race relations and polycultural sensitivity training on college campuses. *Journal of College Student Psychotherapy*, *8*(1-2), 151-176.

Berkowitz, L., & Lutterman, K. G. (1968). The traditional socially responsible personality. *Public Opinion Quarterly*, *32*(2), 169-185.

Bohner, G., Jarvis, C. I., Eyssel, F., & Siebler, F. (2005). The causal impact of rape myth acceptance on men's rape proclivity: Comparing sexually coercive and noncoercive men. *European Journal of Social Psychology*, *35*(6), 819-828.

Bollinger, C., & Mata, J. (2018). Institutional culture and violence. In C. Bollinger, R. Flintoft, J. Nicoletti, S. Spencer-Thomas, & M. Dvoskina (Eds.), *Violence goes to college: The authoritative guide to prevention, intervention, and response* (pp. 32-44). Springfield, IL: Ch*arles C. Thomas Publisher, LTD.*

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Boone, W.J., Staver, J.R., & Yale, M.S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer.

Bourdieu, P. (1990). *The logic of practice*. Stanford, CA: Stanford University Press.

Bowman, N. A. (2009). College diversity courses and cognitive development among students from privileged and marginalized groups. *Journal of Diversity in Higher Education*, *2*(3), 182.

Bowman, N., Felix, V., & Ortis, L. (2014). Religious/worldview identification and college student success. *Religion & Education, 41*(2), 117-133.

Braidottti, R. (1994). *Nomadic subjects: Embodiment and sexual difference in contemporary feminist theory.* New York, NY: Columbia University Press.

Brinkman, B. G., Dean, A. M., Simpson, C. K., McGinley, M., & Rosén, L. A. (2015). Bystander intervention during college: Women's experiences of gender prejudice. *Sex Roles, 72*(11–12), 485–498.

Brody, N., & Vangelisti, A. L. (2016). Bystander intervention in cyberbullying. *Communication Monographs*, *83*(1), 94–119.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Bronfenbrenner, U. (1993). Ecological models of human development. In M. Gauvain & M. Cole (Eds.), *Readings on the development of children* (2nd ed., pp. 37–43). Oxford, UK: Elsevier.

Brown, A. L., Banyard, V. L., & Moynihan, M. M. (2014). College students as helpful bystanders against sexual violence: Gender, race, and year in college moderate the impact of perceived peer norms. *Psychology of Women Quarterly, 38*(3), 350–362.

Brown, V., & DeCoster, D. (1989). The disturbed and disturbing student. In U. Delworth (Ed.), *Dealing with the behavioral and psychological problems ofstudents: New directions for student services* (Vol. 45, pp. 43–56). San Francisco, CA: Jossey-Bass.

Brownmiller, S. (1975). *Against our will: Men, women, and rape*. New York, NY: Simon & Schuster.

Bryan, J. H., & Test, M. A. (1967). Models and helping: naturalistic studies in aiding behavior. *Journal of Personality and Social Psychology*, *6*(4, Pt.1), 400-407.

Bubolz, M. M., & Sontag, S. (1998). Human ecology theory. In P. Boss, W. Doherty, R. LaRossa, W. Schumm, & S. Steinmetz (Eds.), *Sourcebook for family theories and methods: A contextual approach* (pp. 419–450). New York, NY: Plenum.

Buddie, A. M., & Miller, A. G. (2001). Beyond rape myths: A more complex view of perceptions of rape victims. *Sex Roles*, *45*(3-4), 139-160.

Burn, S. M. (2009). A situational model of sexual assault prevention through bystander intervention. *Sex Roles*, *60*(11–12), 779–792.

Burt, M. R. (1980). Cultural myths and support for rape. *Journal of Personality and Social Psychology*, *39*(2), 217–230.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge, UK: Cambridge University Press.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London, UK: Longmans Green

Campus Technical Assistance and Resource Project. (n.d.). Addressing gender-based violence on college campuses: Guide to a comprehensive model. Retrieved from http://changingourcampus.org/documents/FINAL-GBV-Comprehensive-Model-22117.pdf

Cantalupo, N. C. (2009). Campus violence: Understanding the extraordinary through the ordinary. *Journal of College and University Law, 35*, 3, 613-690.

Cantor, D., Fisher, B., Chibnall, S., Townsend, R., Lee, H., Bruce, C., & Thomas, G. (2015). *Report on the AAU Campus Climate Survey on Sexual Assault and Sexual Misconduct*. Washington, DC: Association of American Universities. Retrieved from http://www.aau.edu/ uploadedFiles/AAU_Publications/AAU_Reports/Sexual_Assault_Campus_Survey/Report%20on%20the%20AAU%20Campus%20Climate%20Survey%20on%20Sexual%20Assault%20and%20Sexual%20Misconduct.pdf

Carlo, G., & Randall, B. A. (2001). Are all prosocial behaviors equal? A socioecological developmental conception of prosocial behavior. In F. H. Columbus (Ed.), *Advances in* *psychology research* (Vol. 3, pp. 151–170). Hauppauge, NY: Nova Science Publishers, Inc.

Carlo, G., & Randall, B. A. (2002). The development of a measure of prosocial behaviors for late adolescents. *Journal of Youth and Adolescence, 31*(1), 31–44.

Carlo, G., Eisenberg, N., & Knight, G. P. (1992). An objective measure of adolescents' prosocial moral reasoning. *Journal of Research on Adolescence*, *2*(4), 331-349.

Carlson, M. (2008). I'd rather go along and be considered a man: Masculinity and bystander intervention. *The Journal of Men's Studies*, *16*(1), 3–17.

Carmines, E. G., & R. A. Zeller. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Carr, J. L. (2005). *American College Health Association campus violence white paper*. Baltimore, MD: American College Health Association.

Casey, E. A., & Ohler, K. (2012). Being a positive bystander: Male antiviolence allies' experiences of "stepping up". *Journal of Interpersonal Violence*, *27*(1), 62-83.

Cassidy, L., & Hurrell, R. (1995). The influence of victim's attire on adolescents' judgments of date rape. *Adolescence, 30,* 319–323.

Castello, J., Coomer, C., Stillwell, J., & Cate, K. L. (2006). The attribution of responsibility in acquaintance rape involving ecstasy. *North American Journal of Psychology, 8,* 411–420.

Chekroun, P., & Brauer, M. (2002). The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology, 32*(6), 853-867.

Chickering, A. W., & Reisser, L. (1993). *Education and identity.* San Francisco, CA: Jossey-Bass.

Cho, S., Crenshaw, K. W., & McCall, L. (2013). Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of Women in Culture and Society*, *38*(4), 785-810.

Clark, L. A., & Watson, D. (1999). Temperament: A new paradigm for trait psychology. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 399–423). New York, NY: The Guilford Press.

Clark, R. D., & Word, L. E. (1974). Where is the apathetic bystander? Situational characteristics of the emergency. *Journal of Personality and Social Psychology*, *29*(3), 279-287.

Coker, A. L., Cook-Craig, P. G., Williams, C. M., Fisher, B. S., Clear, E. R., Garcia, L. S., & Hegge, L. M. (2011). Evaluation of Green Dot: An active bystander intervention to reduce sexual violence on college campuses. *Violence Against Women*, *17*(6), 777–796.

Collins, P. H, & Chepp, V. (2013). Intersectionality. In G. Waylen, K. Celis, J. Kanola, & S. L. Weldon (Eds.), *Oxford handbook of gender and politics* (pp. 57-87). New York, NY: Oxford University Press.

Collins, P. H. (2015). Intersectionality's definitional dilemmas. *Annual Review of Sociology, 41*, 1-20.

Comack, E., & Peter, T. (2005) How the criminal justice system responds to sexual assault survivors: The slippage between ‘responsibilization’ and ‘blaming the victim’. C*anadian Journal of Women and the Law, 17*(2), 283–309.

Comack, E., & Peter, T. (2005). How the criminal justice system responds to sexual assault survivors: The slippage between "responsibilization" and "blaming the victim". *Canadian Journal of Women and the Law*, *17*(2), 283-309.

Cox, T. (1991). Study examines reasons for lack of campus diversity. *Black Issues in Higher Education*, *8*,(13), 3-4.

Cramer, R. E., McMaster, M. R., Bartell, P. A., & Dragna, M. (1988). Subject competence and the minimization of the bystander effect. *Journal of Applied Social Psychology, 18,* 1133–1148.

Creamer, E. G., Magolda, M. B., & Yue, J. (2010). Preliminary evidence of the reliability and validity of a quantitative measure of self-authorship. *Journal of College Student Development*, *51*(5), 550-562.

Crenshaw, K. W. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *140*, 139–167.

Crenshaw, K. W. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *46*, 1241–1299.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Ps*ychometrika*, *16*(3), 297-334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York, NY: Wiley.

Dahl, L. S., Lo, M. A., Youngerman, E., & Mayhew, M. J. (2017, March). *Activating the potential for bystander intervention on campus.* Presentation at the annual meeting of ACPA: College Student Educators International, Columbus, OH.

Dahlberg, L. L., & Krug, E. G. (2002). Violence a global public health problem. In E. G. Krug, L. L. Dahlberg, J. A. Mercy, A. B. Zwi, & R. Lozano (Eds.), *The world report on violence and health* (pp. 1–22). Geneva, Switzerland: World Health Organization.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377-383.

David, G. R. (1965). Stating objectives appropriately for program, for curriculum, and for instructional materials development. *Journal of Teacher Education*, *16*(1), 83-92.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113.

De Caroli, M. E., & Sagone, E. (2013). Self-efficacy and prosocial tendencies in Italian adolescents. *Procedia - Social and Behavioral Sciences*, *92*, 239–245.

Dekovic, M., & Janssens, J. M. A. M. (1992). Parents' child-rearing style and child's sociometric status. *Developmental Psychology*, *28*(5), 925–932.

DeVellis, R. F. (2017). *Scale development: Theory and applications (4th ed.)*. Thousand Oaks, CA: Sage.

Dill, B. T., & Zambrana, R. E. (2009). Critical thinking about inequality: An emerging lens. In B. T. Dill & R. E. Zambrana (Eds.), *Emerging intersections: Race, class, and gender in theory, policy, and practice* (pp. 1-21). New Brunswick, NJ: Rutgers University Press.

Donovan, R. A. (2007). To blame or not to blame: Influences of target race and observer sex on rape blame attribution. *Journal of Interpersonal Violence*, *22*(6), 722-736.

Dorough, S. (2011). Moral development. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of child behavior and development* (pp. 967–970). New York, NY: Springer.

Dovidio, J. F. (1984). Helping behavior and altruism: An empirical and conceptual overview. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 17, pp. 361-427). Orlando, FL: Academic Press.

Dovidio, J. F., & Gaertner, S. L. (1983). The effects of sex, status, and ability on helping behavior. *Journal of Applied Social Psychology*, *13*(3), 191-205.

Dovidio, J. F., & Gaertner, S. L. (2004). On the nature of contemporary prejudice. In P. S. Rothenberg (Ed.), *Race, class, and gender in the United States: An integrated study* (pp. 132-142). New York, NY: Worth Publishers.

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. A. (2006). *The social psychology of prosocial behavior*. New York, NY: Psychology Press.

Du Mont, J., Miller, K. L., & Myhr, T. L. (2003). The role of "real rape" and "real victim" stereotypes in the police reporting practices of sexually assaulted women. *Violence Against Women*, *9*(4), 466-486.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York, NY: Russell Sage.

Earley, P. C., & Ang, S. (2003). *Cultural intelligence: Individual interactions across cultures*. Standford, CA: Stanford University Press.

Eisenberg, N. (1986). The development of social knowledge: Morality and convention. *American Scientist, 74*(2), 204-205

Eisenberg, N., & Fabes, R. A. (1998). Prosocial development. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology, Vol. 3: Social, emotional, and personality development* (5th ed., pp. 701-778). New York: John Wiley & Sons.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91–119.

Eisenberg, N., Carlo, G., Murphy, B., & Court, P. (1995). Prosocial development in late adolescence: a longitudinal stu*dy. Child Development*, *66*(4), 1179-1197.

Ethier, K. A., & Deaux, K. (1994). Negotiating social identity when contexts change: Maintaining identification and responding to threat. *Journal of Personality and Social Psychology, 67*(2), 243.

Evans, N. J., Forney, D. S., Guido, F. M., Patton, L. D., & Renn, K. A. (2010). *Student development in college: Theory, research, and practice* (2nd ed.). San Francisco, CA: Jossey-Bass.

Eyssel, F., & Bohner, G. (2011). Schema effects of rape myth acceptance on judgments of guilt and blame in rape cases: The role of perceived entitlement to judge. *Journal of Interpersonal Violence*, *26*(8), 1579-1605.

Federal Bureau of Investigation. (2016). *2015 hate crime statistics*. Washington, DC: U.S. Department of Justice. Retrieved from https://ucr.fbi.gov/hate-crime/2015

Ferrans, S. D., Selman, R. L., & Feigenberg, L. F. (2012). Rules of the culture and personal needs: Witnesses' decision-making processes to deal with situations of bullying in middle school. *Harvard Educational Review, 82*, 445–470.

Field, H. S. (1978). Attitudes toward rape: A comparative analysis of police, rapists, crisis counselors and citizens. *Journal of Personality and Social Psychology, 36*, 156-179.

Fischer, P., Greitemeyer, T., Pollozek, F., & Frey, D. (2006). The unresponsive bystander: Are bystanders more responsive in dangerous emergencies? *European Journal of Social Psychology, 36*(2), 267–278.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., . . . Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517-537.

Fisher, B., Cullen, F. T., & Turner, M. G. (1999). *The extent and nature of the sexual victimization of college women: A national level analysis*. Washington, DC: National Institute of Justice.

Flanders, C. E., Dobinson, C., & Logie, C. (2017). Young bisexual women's perspectives on the relationship between bisexual stigma, mental health, and sexual health: A qualitative study. *Critical Public Health*, *27*(1), 75-85.

Fogle, J. (2000). Acquaintance rape and the attribution of responsibility: The role of alcohol and individual differences. *IU South Bend Undergraduate Research Journal, 3*, 24-29.

Ford, J., & Soto-Marquez, J. G. (2016). Sexual assault victimization among straight, gay/lesbian, and bisexual college students. *Violence and Fender, 3*(2), 107-115.

Foubert, J. D., & Bridges, A. J. (2017a). Predicting bystander efficacy and willingness to intervene in college men and women. *Violence Against Women, 23*(6), 692–706.

Foubert, J. D., & Bridges, A. J. (2017b). What is the attraction? Pornography use motives in relation to bystander intervention. *Journal of Interpersonal Violence, 32*(20), 3071–3089.

Foubert, J. D., Brosi, M. W., & Bannon, R. S. (2011). Pornography viewing among fraternity men: Effects on bystander intervention, rape myth acceptance and behavioral intent to commit sexual assault. *Sexual Addiction and Compulsivity, 18*(4), 212–231.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*, *21*(3), 242-252.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1-13.

Gaertner, S. L., Dovidio, J. F., & Johnson, G. (1982). Race of victim, nonresponsive bystanders, and helping behavior. *The Journal of Social Psychology*, *117*, 69–77.

Garcia, G. A., Johnston, M. P., Garibay, J. C., Herrera, F. A., & Giraldo, L. G. (2011). When parties become racialized: Deconstructing racially themed parties. *Journal of Student Affairs Research and Practice*, *48*(1), 5-21.

Geen, R. G. (1998). Aggression and antisocial behavior. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 317–356). New York, NY: McGraw-Hill.

Gelin, M. N., Beasley, T. M., & Zumbo, B. D. (2003, April). *What is the impact on scale reliability and exploratory factor analysis of a Pearson correlation matrix when some respondents are not able to follow the rating scale?.* Paper presented at the annual meeting of the American Educational Research Association (AERA) in Chicago, Illinois. Retrieved from http://faculty.educ.ubc.ca/zumbo/aera/papers/GelinBeasleyZumbo_7Apr.pdf

George, W. H., & Martínez, L. J. (2002). Victim blaming in rape: Effects of victim and perpetrator race, type of rape, and participant racism. *Psychology of Women Quarterly*, *26*(2), 110-119.

Goldman, J. A., & Harlow, L. L. (1993). Self-perception variables that mediate AIDS-preventive behavior in college students. *Health Psychology*, *12*(6), 489-498.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528.

Gottlieb, J., & Carver, C. S. (1980). Anticipation of future inter action and the bystander effect. *Journal of Experimental Social Psychology, 16*, 253-260

Gregg, R. B. (1966). *The power of nonviolence*. New York, NY: Shocken Books.

Grimley, D., Prochaska, J. O., Velicer, W. F., Vlais, L. M., & DiClemente, C. C. (1994). The transtheoretical model of change. In T.M. Brinthaupt & R.P. Lipka (Eds.), *Changing the self: Philosophies, techniques, and experiences* (p. 201 – 227). Albany, NY: State University of New York Press.

Hancock, A. M. (2016). *Intersectionality: An intellectual history*. New York, NY: Oxford University Press.

Hanson, D., Turbett, P., & Whelehan, P. (1986). Interpersonal violence: Addressing the problem on a college campus. In *Proceedings of University Symposium on Personal Safety,* State University of New York, Albany, NY.

Harari, H., Harari, O., & White, R. V. (1985). The reaction to rape by American male bystanders. *The Journal of Social Psychology*, *125*(5), 653–658.

Harrell, W. A. (1978). Physical attractiveness, self-disclosure, and helping behavior. *The Journal of Social Psychology*, *104*(1), 15-17.

Harris, K. L. (2017). Re-situating organizational knowledge: Violence, intersectionality and the privilege of partial perspective. *Human Relations*, *70*(3), 263-285.

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Chicago, IL: University of Chicago Press.

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science*, *24*(4), 366-374.

Herstein, I. N. (1999). *Abstract algebra*. New York, NY: J. Wiley & Sons.

Hoffman, M. L. (1970). Moral development. *Carmichael's Manual of Child Psychology*, *2*, 261-359.

Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge, UK: Cambridge University Press.

Hong, L. (2017). Digging up the roots, rustling the leaves: A critical consideration of the root causes of sexual violence and why higher education needs more courage. In J. C. Harris & C. Linder (Eds.), *Intersections of identity and sexual violence on campus: Centering minoritized students experiences* (pp. 23–41). Sterling, VA: Stylus.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation*. Needham Heights, MA: Allyn & Bacon.

Howard, W., & Crano, W. D. (1974). Effects of sex, conversation, location, and size of observer group on bystander intervention in a high risk situation. *Sociometry, 37,* 491–507.

Hoxmeier, J. C., Acock, A. C., & Flay, B. R. (2017). Students as prosocial bystanders to sexual assault: Demographic correlates of intervention norms, intentions, and missed opportunities. *Journal of Interpersonal Violence*, 1–24.

Iverson, S. V. (2017). Mapping identities: An intersectional analysis of policies on sexual violence. In J. C. Harris & C. Linder (Eds.), *Intersections of identity and sexual violence on campus: Centering minoritized students experiences* (pp. 214-234). Sterling, VA: Stylus.

Johnson, A. G. (2006). *Privilege, power, and difference* (2nd ed.). New York, NY: McGraw-Hill.

Johnson, J. (2018, April 16). From Starbucks to hashtags: We need to talk about why white Americans call the police on Black people [Blog post]. Retrieved from https://www.theroot.com/from-starbucks-to-hashtags-we-need-to-talk-about-why-w-1825284087

Johnson, R. C., Danko, G. P., Darvill, T. J., Bochner, S., Bowers, J. K., Huang, Y. H., ... & Pennington, D. (1989). Cross-cultural assessment of altruism and its correlates. *Personality and Individual Differences, 10*(8), 855-868.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 485–498). Washington, DC: American Council on Education.

Jones, S. R., & Abes, E. S. (2013). *Identity development of college students: Advancing frameworks for multiple dimensions of identity*. San Francisco, CA: Jossey-Bass.

Jones, S. R., & McEwen, M. K. (2000). A conceptual model of multiple dimensions of identity. *Journal of College Student Development, 41*(4), 405-414.

Jorm, A. F., Korten, A. E., Rodgers, B., Jacomb, P. A., & Christensen, H. (2002). Sexual orientation and mental health: results from a community survey of young and middle–aged adults. *The British Journal of Psychiatry, 180*(5), 423-427.

Katz, J., Merrilees, C., Hoxmeier, J. C., & Motisi, M. (2017). White female bystanders' responses to a Black woman at risk for incapacitated sexual assault. *Psychology of Women Quarterly, 41*(2), 273-285.

Katz, J., Merrilees, C., LaRose, J., & Edgington, C. (2018). White female bystanders' responses to a Black woman at risk for sexual assault: Associations with attitudes about sexism and racial injustice. *Journal of Aggression, Maltreatment & Trauma, 27*(4), 444-459

Kegan, R. (1982). *The evolving self*. Cambridge, MA: Harvard University Press.

Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Cambridge, MA: Harvard University Press.

Kelley, K., & Byrne, D. (1976). Attraction and altruism: With a little help from my friends. *Journal of Research in Personality*, *10*(1), 59-68.

Kelly, J. R., & McGrath, J. E. (1988). *On time and method.* Newbury Park, CA: Sage

Kelly, L. (1987). The continuum of sexual violence. In Hanmer, J. & Maynard, M. (Eds.), *Women, violence and social control* (pp. 46-60). Atlantic Highlands, NJ: Humanities Press International.

Kelly, L. (1989). *Surviving sexual violence*. Minneapolis, MN: University of Minnesota Press.

Kerlinger, F. N. (1973). *Foundations of behavioral science* (2nd ed.). New York, NY: Holt, Renehard and Winston.

Kilpatrick, D. G., Resnick, H. S., Ruggiero, K. J., Conoscenti, L. M., & McCauley, J. (2007). *Drug-facilitated, incapacitated, and forcible rape: A national study* (Final report submitted to the National Institute of Justice [NCJ 219181]). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Kitzrow, M. A. (2003). The mental health needs of today's college students: Challenges and recommendations. *NASPA Journal*, *41*(1), 167-181.

Kleinke, C. L. (1977). Effects of dress on compliance to requests in a field setting. *The Journal of Social Psychology, 101*(2), 223-224.

Korman, A. T., & Greenstein, S. (2016). *The Culture of Respect CORE Blueprint*. Washington, DC: National Association of Student Personnel Administrators (NASPA). Retrieved from http://archive.naspa.org/files/NASPA_CoR_PilotProgramReport_FINAL.pdf

Korman, A. T., Greenstein, S., Wesaw, A., & Hopp, J. (2017). *Institutional responses to sexual violence: What data from a culture of respect program tell us about the state of the field*. Washington, DC: National Association of Student Personnel Administrators (NASPA). Retrieved from https://www.naspa.org/images/uploads/main/CultureofRespectLandscapeReportFINAL.pdf

Kosmin, B. A., & Keysar, A. (2015). *National demographic survey of American Jewish college students 2014: Anti-Semitism report.* Hartford, CT: The Louis D. Brandeis Center, Trinity College.

Koss, M. P., Gidycz, C. A., & Wisniewski, N. (1987). The scope of rape: Incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. *Journal of Consulting and Clinical Psychology, 55*(2), 162.

Kou, Y., Hong, H. F., Tan, C., & Li, L. (2007). Revisioning prosocial tendencies measure for adolescents. *Psychological Development and Education, 23*, 112–117.

Krebs, C. P., Lindquist, C. H., Warner, T. D., Fisher, B. S., & Martin, S. L. (2007). *The campus sexual assault (CSA) study: Final report*. Washington, DC: National Institute of Justice, US Department of Justice.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236.

Kunstman, J. W., & Plant, E. A. (2008). Racing to help: Racial bias in high emergency helping situations. *Journal of Personality and Social Psychology*, *95*(6), 1499–1510.

Lanier, C. A., & Elliot, M. N. (1997). A new instrument for the evaluation of a date rape prevention program. *Journal of College Student Development, 38(*6), 673–676.

Lanier, C. A., & Green, B. A. (2006). Principal component analysis of the College Date Rape Attitude Survey (CDRAS): An instrument for the evaluation of date rape prevention programs. *Journal of Aggression, Maltreatment & Trauma, 13*(2), 79–93.

LaPlant, L. E. (2002). *Implementation and evaluation of group-based prevention of eating concerns using self-efficacy and knowledge enhancement* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses. (Order No. 3045332).

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York, NY: Appleton-Century-Crofts.

Leidig, M. J. (1992). The continuum of violence against women: Psychological and physical consequences. *Journal of American College Health, 40*, 149-155.

Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self categorization and bystander non-intervention. *Journal of Applied Social Psychology, 32*, 1452–1463.

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

Li, T., Dobinson, C., Scheim, A. I., & Ross, L. E. (2013). Unique issues bisexual people face in intimate relationships: A descriptive exploration of lived experience. *Journal of Gay & Lesbian Mental Health, 17*(1), 21-39.

Likert, R. (1932). *A technique for the measurement of attitudes*. New York, NY: The Science Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7*(1), 129-139.

Linacre, J. M. (2018). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago IL: Winsteps.com.

Linacre, J. M. (n.d.). *Dimensionality: Contrasts & variances*. Retrieved from https://www.winsteps.com/winman/principalcomponents.htm

Linder, C. (2018). *Sexual violence on campus: Power-conscious approaches to awareness, prevention, and response*. Bingley, UK: Emerald Publishing Limited

Linder, C., & Harris, J. C. (2017). Conclusion: History, identity, and power-conscious strategies for addressing sexual violence on college campuses. In J. C. Harris & C. Linder (Eds.), *Intersections of identity and sexual violence on campus: Centering minoritized students experiences* (pp. 235–256). Sterling, VA: Stylus.

Loewenstein, G., & Small, D. A. (2007). The Scarecrow and the Tin Man: The vicissitudes of human sympathy and caring. *Review of General Psychology*, *11*(2), 112-126.

Lonsway, K. A., & Fitzgerald, L. F. (1994). Rape myths: In review. *Psychology of Women Quarterly*, *18*, 133–164.

Lonsway, K.A. & Kothari, C. (2000). First year campus acquaintance rape education. *Psychology of Women Quarterly, 24*, 220-232.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores.* Charlotte, NC: Information Age.

Love, P. G., & Guthrie, V. L. (eds.) (1999). Understanding and applying cognitive development theory. *New Directions for Student Services* (No. 88), San Francisco, CA: Jossey-Bass.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement, 55*(6), 967-975.

Malamuth, N. M., Sockloskie, R. J., Koss, M. P., & Tanaka, J. S. (1991). Characteristics of aggressors against women: Testing a model using a national sample of college students. *Journal of Consulting and Clinical Psychology*, *59*(5), 670.

Manning, R., Levine, M., & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist*, *62*(6), 555.

March, D. J., & Olson, J. (1980). *Ambiguity and choice in organizations*. New York, NY: Oxford University Press.

Martin, P. Y., & Hummer, R. A. (1989). Fraternities and rape on campus. *Gender & Society*, *3*(4), 457-473.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*(2), 157-176.

Mayhew, M. J., & Fernández, S. D. (2007). Pedagogical practices that contribute to social justice outcomes. *The Review of Higher Education, 31*(1), 55-80.

Mayhew, M. J., Caldwell, R. J., & Goldman, E. G. (2011). Defining campus violence: A phenomenological analysis of community stakeholder perspectives. *Journal of College Student Development*, *52*(3), 253-269.

Mayhew, M. J., Lo, M. A., Dahl, L. S., & Selznick, B. S. (2018). Assessing students' intention to intervene in a bystander situation. *Journal of College Student Development*, *59*(6), 762-768.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Lo, M. A., Starcke, M. A., Riggers-Piehl, T., & Crandall, R. E. (2017). Expanding perspectives on evangelicalism: How non-evangelical students appreciate evangelical Christianity. *Review of Religious Research, 59*(2), 207-230.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Seifert, T. A., Wolniak, G. C., Pascarella, E. T., & Terenzini, P. T. (2016). *How college affects students, volume 3: 21st century evidence that higher education works*. San Francisco, CA: Jossey-Bass.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: L. Erlbaum Associates.

McIntosh, P. (1990). White privilege: Unpacking the invisible knapsack. *Independent School, 49*(2), 31–35.

McMahon, S. (2010). Rape myth beliefs and bystander attitudes among incoming college students. *Journal of American College Health*, *59*(1), 3–11.

McMahon, S., & Banyard, V. L. (2012). When can I help? A conceptual framework for the prevention of sexual violence through bystander intervention. *Trauma, Violence, & Abuse*, *13*(1), 3–14.

McMahon, S., & Dick, A. (2011). “Being in a room with like-minded men”: An exploratory study of men’s participation in a bystander intervention program to prevent intimate partner violence. *The Journal of Men’s Studies, 19*(1), 3–18.

McMahon, S., Allen, C. T., Postmus, J. L., McMahon, S. M., Peterson, N. A., & Hoffman, M. L. (2014). Measuring bystander attitudes and behavior to prevent sexual violence. *Journal of American College Health*, *62*(1), 58–66.

McMahon, S., Palmer, J. E., Banyard, V., Murphy, M., & Gidycz, C. A. (2017). Measuring bystander behavior in the context of sexual violence prevention: Lessons learned and new directions. *Journal of Interpersonal Violence, 32*(16), 2396–2418.

McMahon, S., Postmus, J. L., & Koenick, R. A. (2011). Conceptualizing the engaging bystander approach to sexual violence prevention on college campuses. *Journal of College Student Development, 52*(1), 115–130.

Mead, M. (1969). Violence and its regulation: How do children learn to govern their own violent impulses. *American Journal of Orthopsychiatry, 39*(2), 227-229.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Messman-Moore, T. L., Coates, A. A., Gaffey, K. J., & Johnson, C. F. (2008). Sexuality, substance use, and susceptibility to victimization: Risk for rape and sexual coercion in a prospective study of college women. *Journal of Interpersonal Violence, 23*(12), 1730-1746.

Molina, B. (2018, December 20). Cashing checks, napping, more activities leading to police calls on black people in 2018. *USA Today*. Retrieved from https://www.usatoday.com/story/news/nation/2018/12/20/black-people-doing-normal-things-who-had-police-called-them-2018/2374750002/

Moradi, B., & Grzanka, P. R. (2017). Using intersectionality responsibly: Toward critical epistemology, structural analysis, and social justice activism. *Journal of Counseling Psychology*, *64*(5), 500-513.

Museus, S. D., & Griffin, K. A. (2011). Mapping the margins in higher education. *New Directions for Institutional Research*, *151*, 5–13.

National Association of Student Personnel Administrators [NASPA]. (1989). *Preliminary report: Task group on campus safety and security* (Brochure). Washington, D.C.: Author.

Nelson, J. K., Dunn, K. M., & Paradies, Y. (2011). Bystander anti-racism: A review of the literature. *Analyses of Social Issues and Public Policy*, *11*(1), 263–284.

Ngai, S. S., & Xie, L. (2018). Toward a validation of the prosocial tendencies measure among Chinese adolescents in Hong Kong. *Child Indicators Research, 11*(4), 1281-1299.

Nicholson, M. E., Maney, D. W., Blair, K., Wamboldt, P. M., Mahoney, B. V., & Yuan, J. (1998). Trends in alcohol related campus violence: Implications for prevention. *Journal of Alcohol and Drug Prevention, 43*(3), 34-52.

Nicksa, S. C. (2014). Bystander's willingness to report theft, physical assault, and sexual assault. *Journal of Interpersonal Violence, 29*(2), 217–236.

Nicoletti, J., Spencer-Thomas, S., & Dvoskina, M. (2018). Threats of violence. In C. Bollinger, R. Flintoft, J. Nicoletti, S. Spencer-Thomas, & M. Dvoskina (Eds.), *Violence goes to college: The authoritative guide to prevention, intervention, and response* (pp. 45-60). Springfield, IL: Charles C. Thomas Publisher, LTD.

Noori Farzan, A. (2018, October 19). BBQ Becky, Permit Patty and Cornerstore Caroline: Too 'cutesy' for those white women calling police on black people?. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2018/10/19/bbq-becky-permit-patty-and-cornerstore-caroline-too-cutesy-for-those-white-women-calling-cops-on-blacks

Norris, J., & Cubbins, L. A. (1992). Effects of victims' and assailants' alcohol consumption on judgments of their behavior and traits. *Psychology of Women Quarterly, 16*, 179–191.

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2nd ed.). New York, NY: McGraw-Hill

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

O'Donnell, N. (2018, October 16). #LivingWhileBlack: Videos document Black people being harassed as they golf, swim, eat, nap. *NBC Connecticut*. Retrieved from https://www.nbcconnecticut.com/news/national-international/Living-While-Black-497623161.html

O'Donohue, W., Yeater, E. A., & Fanetti, M. (2003). Rape prevention with college males: The roles of rape myth acceptance, victim empathy, and outcome expectancies. *Journal of Interpersonal Violence*, *18*(5), 513-531.

Osborne, R. (1995). The continuum of violence against women in Canadian universities: Toward a new understanding of the chilly campus climate. *Women's Studies International Forum, 18,* 637-646.

Paul, J. P. (1984). The bisexual identity: An idea without social recognition. *Journal of Homosexuality, 9*(2-3), 45-63.

Payne, B. K. (2008). Challenges responding to sexual violence: Differences between college campuses and communities. *Journal of Criminal Justice, 36*(3), 224-230.

Payne, D. L., Lonsway, K. A., & Fitzgerald, L. F. (1999). Rape myth acceptance: Exploration of its structure and its measurement using the Illinois Rape Myth Acceptance Scale. *Journal of Research in Personality, 33*, 27-68.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated analysis. *Journal of Econometrics*, *22*, 229-243.

Perry, W. G., Jr. (1968). *Patterns of development in thought and values of students in a liberal arts college: A validation of a scheme.* Cambridge, MA: Bureau of Study Counsel, Harvard University.

Pezza, P. E., & Bellotti, A. (1995). College campus violence: Origins, impacts, and responses. *Educational Psychology Review*, *7*(1), 105-123.

Piliavin, J. A., Dovidio, J., Gaertner, S., & Clark, R.D., III. (1981). *Emergency intervention*. New York, NY: Academic Press.

Pinzone-Glover, H. A., Gidycz, C. A., & Jacobs, C. D. (1998). An acquaintance rape prevention program: Effects on attitudes toward women, rape-related attitudes, and perceptions of rape scenarios. *Psychology of Women Quarterly*, *22*, 605-621.

Pizzolato, J. E. (2007). Assessing self-authorship. *New Directions for Teaching and Learning, 2007*(109), 31-42.

Pozzoli, T., Gini, G., & Vieno, A. (2012). The role of individual correlates and class norms in defending and passive bystanding behavior in bullying: A multilevel analysis. *Child Development, 83*, 1917–1931.

Presley, C. A., Meilman, P. W., & Cashin, J. R. (1997). Weapon carrying and substance abuse among college students. *Journal of American College Health*, *46*(1), 3-8.

Rand, M. R. (2009). *Criminal victimization, 2008*. Washington, DC, Bureau of Justice Statistics, US Department of Justice. Retrieved from https://www.bjs.gov/content/pub/pdf/cv08.pdf

Rankin, S. (2003). *Campus climate for sexual minorities: A national perspective.* New York: NY: National Gay and Lesbian Task Force Policy Institute.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Ray, C. M. (2007). *Development of an integrated model and measure of the moral dimensions of justice and care* (Unpublished doctoral dissertation). Oklahoma State University, Stillwater, OK. Retrieved from Proquest Dissertations and Theses.

Renn, K. A., & Arnold, K. D. (2003). Reconceptualizing research on college student peer culture. *The Journal of Higher Education*, *74*(3), 261–291.

Revelle W. (1979) Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57–74

Riggins, S. (1997). The rhetoric of othering. In S. Riggins (ed.) *The language and politics of exclusion: Others in discourse* (pp. 1–31). Thousand Oaks, CA: Sage.

Roark, M. L. (1987). Preventing violence on college campuses. *Journal of Counseling & Development*, *65*(7), 367-371.

Roark, M. L. (1994). Conceptualizing campus violence: Definitions, underlying factors, and effects. *Journal of College Student Psychotherapy*, *8*(1-2), 1-28.

Rodrigues, J., Ulrich, N., Mussel, P., Carlo, G., & Hewig, J. (2017). Measuring prosocial tendencies in Germany: Sources of validity and reliablity of the revised prosocial tendency measure. *Frontiers in Psychology, 8*, 1–17.

Rushton, J. P., & Campbell, A. C. (1977). Modeling, vicarious reinforcement and extraversion on blood donating in adults: Immediate and long-term effects. *European Journal of Social Psychology, 7*(3), 297-306.

Rushton, J. P., Chrisjohn, R. D., & Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and Individual Differences, 2*(4), 293-302.

Sampson, R. (2003). *Acquaintance rape of college students.* U.S. Department of Justice: Office of Community Oriented Policing Services.

Saucier, D. A., Miller, C. T., & Doucet, N. (2005). Differences in helping whites and Blacks: A meta-analysis. *Personality and Social Psychology Review*, *9*(1), 2-16.

Saxe, L., Sasson, T., Wright, G., & Hecht, S. (2015). *Antisemitism on the college campus: Perceptions and realities.* Waltham, MA: Cohen Center for Modern Jewish Studies, Brandeis University.

Schult, D. G., & Schneider, L. J. (1991). The role of sexual provocativeness, rape history, and observer gender in perceptions of blame in sexual assault. *Journal of Interpersonal Violence, 6*, 94–101.

Schwartz, S. H. (1968). Words, deeds and the perception of consequences and responsibility in action situations. *Journal of Personality and Social Psychology, 10*(3), 232-242.

Schwartz, S. H., & Gottlieb, A. (1980). Bystander anonymity and reactions to emergencies. *Journal of Personality and Social Psychology*, *39*(3), 418–430.

Schwendinger, J. R., & Schwendinger, H. (1974). Rape myths: In legal, theoretical, and everyday practice. *Crime and Social Justice, 1*, 18-26.

Shotland, R. L., & Heinold, W. D. (1985). Bystander response to arterial bleeding: helping skills, the decision-making process, and differentiating the helping response. *Journal of Personality and Social Psychology*, *49*(2), 347-356.

Shultz, E., Heilman, R., & Hart, K. J. (2014). Cyber-bullying: An exploration of bystander behavior and motivation. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *8*(4), article 3.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika*, *74*(1), 107.

Slaby, R. G., Wilson-Brewer, R., & DeVos, E. (1994). *Agressors, victims, & bystanders: An assessment-based middle school violence prevention curriculum*. Newton, MA: Education Development Center.

Slater, B. R. (1994). Violence against lesbian and gay male college students. *Journal of College Student Psychotherapy*, *8*(1-2), 177-202.

Sloan III, J. J., Fisher, B. S., & Cullen, F. T. (1997). Assessing the student right-to-know and Campus Security Act of 1990: An analysis of the victim reporting practices of college and university students. *Crime & Delinquency*, *43*(2), 148-168.

Solomon, L. Z., Solomon, H., & Maiorca, J. (1982). The effects of bystander’s anonymity, situational ambiguity, and victim’s status on helping. *The Journal of Social Psychology*, *117*(2), 285–294.

Staub, E. (1979). *Positive social behavior and morality: Socialization and development* (Vol. 2). New York, NY: Academic Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). Oxford, England: Wiley.

Stout, K. D. (1991). A continuum of male controls and violence against women: A teaching model. *Journal of Social Work, 27*, 305-320.

Stout, K. D., & McPhail, B. (1998). *Confronting sexism and violence against women: A challenge for social work*. New York, NY: Longman.

Stueve, A., Dash, K., O'Donnell, L., Tehranifar, P., Wilson-Simmons, R., Slaby, R. G., & Link, B. G. (2006). Rethinking the bystander role in school violence prevention. *Health Promotion Practice*, *7*(1), 117–124.

Sturgis, P., Roberts, C., & Smith, P. (2014). Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"?. *Sociological Methods & Research*, *43*(1), 15-38.

Swearer, S. M., & Espelage, D. L. (2004). Introduction: A social-ecological framework of bullying among youth. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 1–12). Mahwah, NJ: Lawrence Erlbaum Associates.

Swisher, J. D., Shute, R. E., & Bibeau, D. (1984). Assessing drug and alcohol abuse: An instrument for planning and evaluation. *Measurement and Evaluation in Counseling and Development*, *17*(2), 91-97.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*(1), 1-39.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529-554.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley

Torres, V., Jones, S. R., & Renn, K. A. (2009). Identity development theories in student affairs: Origins, current status, and new approaches. *Journal of College Student Development*, *50*(6), 577-596.

Twemlow, S. W., & Sacco, F. C. (2013). How and why does bystanding have such a startling impact on the architecture of school bullying and violence? *International Journal of Applied Psychoanalytic Studies, 10*, 289–306.

Tyler, R. W. (1973). Assessing educational achievement in the affective domain. *Measurement in Education, 4*(3), 1–8.

U.S. Department of Justice. (2015). *FY 2012 OVW grant awards by program.* Retrieved from www.ovw.usdoj.gov/fy2012-grant-program.htm#2

Underwood, B., & Moore, B. (1982). Perspective-taking and altruism. *Psychological Bulletin, 91*(1), 143–173.

US Department of Education [US DOE]. (2006). Confronting violence head-on. *Catalyst: A Publication of the U.S. Department of Education's Higher Education Center for Alcohol and Other Drug Abuse and Violence Prevention, 7*(3), 1-2. Retrieved from https://permanent.access.gpo.gov/LPS107142/2006/v7_n3_sp2006.pdf

US Department of Education [US DOE]. (2018). [Online trend data generator]. Hate crimes data trend results. Retrieved from https://ope.ed.gov/campussafety/Trend/public/#/subjects

Van Dyne, L., Ang, S., Ng, K. Y., Rockstuhl, T., Tan, M. L., & Koh, C. (2012). Sub-dimensions of the four factor model of cultural intelligence: Expanding the conceptualization and measurement of cultural intelligence. *Social and Personality Psychology Compass, 6*(4), 295-313.

Walters, J., McKellar, A., Lipton, M., & Karme, L. (1981). What are the pros and cons of coed dorms?. *Medical Aspects of Human Sexuality, 15*(8), 48–56.

Weininger, E.B. (2002). Pierre Bourdieu on social class and symbolic violence. In E. O. Wright (Ed.), *Alternative foundations of class analysis* (pp.119-171). Retreived from https://www.ssc.wisc.edu/~wright/Found-all.pdf

Wessler, S., & Moss, M. (2001). *Hate crimes on campus: The problem and efforts to confront it*. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Whatley, M. A. (2005). The effect of participant sex, victim dress, and traditional attitudes on causal judgments for marital rape and victims. *Journal of Family Violence, 20,* 191–200.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly*, *249*(3), 29-38.

Wing Sue, D. (2017). Microaggressions and "evidence": Empirical or experiential reality?. *Perspectives on Psychological Science*, *12*(1), 170-172.

Wolfe, E. W., & Smith, J. E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II - Validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Workman, J. E., & Freeburn, E. W. (1999). An examination of date rape, victim dress, and perceiver variables within the context of attribution theory. *Sex Roles, 41*, 261–278.

World Health Organization [WHO] Global Consultation on Violence and Health. (1996). *Violence: A public health priority*. Geneva, Switzerland: World Health Organization.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Pres

Wright, B.D. (1967). *Sample-free test calibration and person measurement*. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service. Retrieved from https://www.rasch.org/memo1.htm

Yu, C. H. (2005). Test–retest reliability. In K. Kempf-Leonard (Ed.). *Encyclopedia of social measurement* (Vol. 3, pp. 777–784). San Diego, CA: Academic Press.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. P*sychometrika*, *70*(1), 123-133.

Zoccola, P. M., Green, M. C., Karoutsos, E., Katona, S. M., & Sabini, J. (2011). The embarrassed bystander: Embarrassability and the inhibition of helping. *Personality and Individual Differences*, *51*(8), 925–929.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modelling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21-29.

Appendix A: Bystander Intervention Items

Scenario S1: Incapacitated sexual assault at a party

In reading the prompts below, place yourself in the scenario:

You are at a party when you notice that a male student and a female student are leaving together. He has not been drinking very much. She has been drinking and is clearly very intoxicated. <u>You are a friend of the man</u>.

How likely are you to:

a. Say something to the man (your friend) at the time?
b. Say something to the women (who you don't know) at the time?
c. Say something to the man (your friend) at a later time?
d. Get other people to support you in intervening?
e. Call an authority figure (party host, police, etc.) to intervene?

Scenario S2: Incapacitated sexual assault at a party

In reading the prompts below, place yourself in the scenario:

You are at a party when you notice that a male student and a female student are leaving together. He has not been drinking very much. She has been drinking and is clearly very intoxicated. <u>You are a friend of the woman</u>.

How likely are you to:

a. Say something to the man (who you don't know) at the time?
b. Say something to the women (your friend) at the time?
c. Say something to the women (your friend) at a later time?
d. Get other people to support you in intervening?
e. Call an authority figure (party host, police, etc.) to intervene?

Scenario S3: Incapacitated sexual assault at a party

In reading the prompts below, place yourself in the scenario:

You are at a party when you notice that a male student and a female student are leaving together. He has not been drinking very much. She has been drinking and is clearly very intoxicated. <u>You don’t know either person</u>.
How likely are you to:

a. Say something to the man at the time?
b. Say something to the women at the time?
c. Get other people to support you in intervening?
d. Call an authority figure (party host, police, etc.) to intervene?

Scenario S4: Domestic violence

In reading the prompts below, place yourself in the scenario:

Two students you know who are in a dating relationship live in the apartment/room above you. You hear them arguing often, especially when they have been drinking. You hear noises that lead you to believe that their fighting is now physical. <u>You don't know either person very well</u>.

How likely are you to:

a. Say something to the aggressive student at the time?
b. Say something to the nonaggressive student at the time?
c. Say something to the aggressive student at a later time?
d. Say something to the nonaggressive student at a later time?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, police, apartment manager, etc.) to intervene?

Scenario A1: Racial incident in residence hall

In reading the prompts below, place yourself in the scenario:

You are in common area on your floor and you observe a floor mate you don't know very well writing something on a board and laughing. After *they* leave, you notice that *they* wrote a racial slur directed at another peer with whom you are friends.

How likely are you to:

a. Say something at the time to the student writing the comment?
b. Say something at the time to the peer who was targeted?
c. Say something at a later time to the student writing the comment?
d. Say something at a later time to the peer who was targeted?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, police, residence hall/apartment manager, etc.) to intervene?
g. Remove what they wrote?

Scenario A2: Racial incident in residence hall

In reading the prompts below, place yourself in the scenario:

You are in common area on your floor and you observe a floor mate you don't know very well writing something on a board and laughing. After *they* leave, you notice that *they* wrote a racial slur directed at another peer who you also don’t know very well.

How likely are you to:

a. Say something at the time to the student writing the comment?
b. Say something at the time to the peer who was targeted?
c. Say something at a later time to the student writing the comment?
d. Say something at a later time to the peer who was targeted?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, police, residence hall/apartment manager, etc.) to intervene?
g. Remove what they wrote?

Scenario B1: Stolen bag in the library

In reading the prompts below, place yourself in the scenario:

You are studying at the library when a student sitting alone at a nearby table gets up. *They* left *their* belongings at the table and while *they* are gone, you notice someone you know rummaging through *their* stuff. You don't know the person who left *their* belongings.

How likely are you to:

a. Say something at the time to the student rummaging through the stuff?
b. Say something at a later time to the student rummaging through the stuff?
c. Say something at a later time to the student who left their stuff?
d. Get other people to support you in intervening?
e. Call an authority figure (library staff, police, etc.) to intervene?

Scenario B2: Stolen bag in the library

In reading the prompts below, place yourself in the scenario:

You are studying at the library when a student sitting alone at a nearby table gets up. *They* left *their* belongings at the table and while *they* are gone, you notice another student rummaging through *their* stuff. You don't know either person.

How likely are you to:

a. Say something at the time to the student rummaging through the stuff?
b. Say something at a later time to the student who left their stuff?
c. Get other people to support you in intervening
d. Call an authority figure (library staff, police, etc.) to intervene?

Scenario C1: Religious microaggression in class

In reading the prompts below, place yourself in the scenario:

This semester, you are enrolled in a course you really like with a great professor you don't know very well. One day during lecture, your professor makes a few jokes based on religious stereotypes. You have several friends in the class.

How likely are you to:

a. Say something at the time to your professor in front of the class?
b. Say something at a later time to your professor in private?
c. Get other people to support you in intervening
d. Say something at a later time to an authority figure (another professor, campus staff, etc.)?

Scenario C2: Religious microaggression in class

In reading the prompts below, place yourself in the scenario:

This semester, you are enrolled in a course you really like with a great professor you don't know very well. One day during lecture, your professor makes a few jokes based on religious stereotypes. You don't know anyone in the class very well.

How likely are you to:

a. Say something at the time to your professor in front of the class?
b. Say something at a later time to your professor in private?
c. Get other people to support you in intervening
d. Say something at a later time to an authority figure (another professor, campus staff, etc.)?

Scenario D1: Roommates filming sexual encounter

In reading the prompts below, place yourself in the scenario:

You are in your dorm room watching TV when you get a message in the unofficial group chat for the floor. It says everyone is meeting in a room down the hall to watch a crazy video. When you arrive, you see that a floor mate is livestreaming a video of their roommate making out with another student. You are friends with the student showing the video, but you don't know the students in the video or anyone else watching the video very well.

How likely are you to:

a. Say something at the time to the student showing the video?
b. Say something at the time to the students in the video?
c. Say something at a later time to the student showing the video?
d. Say something at a later time to the students in the video?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, residence hall, etc.) to intervene?

Scenario D2: Roommates filming sexual encounter

In reading the prompts below, place yourself in the scenario:

You are in your dorm room watching TV when you get a message in the unofficial group chat for the floor. It says everyone is meeting in a room down the hall to watch a crazy video. When you arrive, you see that a floor mate is livestreaming a video of their roommate making out with another student. You are friends with the student in the livestream, but you don't know the student showing the video or anyone else watching the video very well.

How likely are you to:

a. Say something at the time to the student showing the video?
b. Say something at the time to your friend (the student in the video)?
c. Say something at a later time to the student showing the video?
d. Say something at a later time to your friend (the student in the video)?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, residence hall, etc.) to intervene?

Scenario D3: Roommates filming sexual encounter

In reading the prompts below, place yourself in the scenario:

You are in your dorm room watching TV when you get a message in the unofficial group chat for the floor. It says everyone is meeting in a room down the hall to watch a crazy video. When you arrive, you see that a floor mate is livestreaming a video of their roommate making out with another student. You are friends with the other students watching the livestream, but you don't know the student showing the video or the students in the video very well.

How likely are you to:

a. Say something at the time to the student showing the video?
b. Say something at the time to the students in the video?
c. Say something at a later time to the student showing the video?
d. Say something at a later time to the students in the video?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, residence hall, etc.) to intervene?

Scenario D4: Roommates filming sexual encounter

In reading the prompts below, place yourself in the scenario:

You are in your dorm room watching TV when you get a message in the unofficial group chat for the floor. It says everyone is meeting in a room down the hall to watch a crazy video. When you arrive, you see that a floor mate is livestreaming a video of their roommate making out with another student. You don't know the student showing the livestream, the students in the video, or the anyone else watching the video very well.

How likely are you to:

a. Say something at the time to the student showing the video?
b. Say something at the time to the students in the video?
c. Say something at a later time to the student showing the video?
d. Say something at a later time to the students in the video?
e. Get other people to support you in intervening?
f. Call an authority figure (RA, residence hall, etc.) to intervene?

Scenario E1: Student organization social media video

In reading the prompts below, place yourself in the scenario:

You are browsing through Instagram when you notice a post by a leader of your student organization. The video consists of *them* saying a bunch of racial slurs. You are not a leader of this organization.

How likely are you to:

a. Say something online at the time to your friend who posted the video?
b. Say something at a later time to your friend who posted the video?
c. Get other people to support you in intervening?
d. Get another leader of your student organization to intervene?

Scenario E2: Student organization social media video

In reading the prompts below, place yourself in the scenario:

You are browsing through Instagram when you notice a post by a leader of your student organization. The video consists of *them* saying a bunch of racial slurs. You are also a leader of this organization.

How likely are you to:

a. Say something online at the time to the student who posted the video?
b. Say something at a later time to the student who posted the video?
c. Get other people to support you in intervening?
d. Get another leader of your student organization to intervene?

Appendix B: Wright-Andrich Map

```
MEASURE    PERSON - MAP - ITEM - Andrich thresholds (modal categories if ordered)
               <more>|<rare>
    5             .  +
                  .  |
                  .  |
                     |
                     |
                  .  |
                     |
    4             .  +
                  .  |
                     |
                  .  |
                  .  |
                  .  |
                  .  |
    3             .  +
                  .  |
                  .  |
                 .#  |                                c2_a  .5
                  . T|                                c1_a  .5
                  .  |
                 .#  |
    2            .#  +                                c1_b  .5
                                                      c2_b  .5
                                                      s4_a  .5
                .##  |                                c2_c  .5
                                                      s1_e  .5
                                                      s4_c  .5
                 .#  |                                c1_c  .5
                                                      c1_d  .5
                                                      c2_d  .5
                                                      s3_a  .5
                                                      s3_d  .5
               .### S|                      c1_a  .4  b1_e  .5
                                            c2_a  .4  d1_f  .5
                                                      d4_b  .5
                                                      s2_e  .5
               .###  |                                a2_c  .5
                                                      b2_d  .5
                                                      d1_b  .5
                                                      d2_f  .5
                                                      d3_b  .5
                                                      d3_f  .5
                                                      d4_c  .5
```

```
                                                  d4_d  .5
                                                  s3_b  .5
                                                  s4_b  .5
         .#####  |                                a2_a  .5
                                                  a2_b  .5
                                                  b1_d  .5
                                                  b2_c  .5
                                                  d1_d  .5
                                                  d1_e  .5
                                                  d3_c  .5
                                                  d3_e  .5
                                                  d4_a  .5
                                                  d4_e  .5
                                                  d4_f  .5
                                                  e1_a  .5
                                                  s3_c  .5
        .######  |                                a1_c  .5
                                                  a2_d  .5
                                                  d3_a  .5
                                                  d3_d  .5
                                                  s1_b  .5
                                                  s4_d  .5
1      .#######  +T           c2_a  .3  c1_b  .4  a1_b  .5
                                        c2_b  .4  a1_e  .5
                                        s1_e  .4  a1_f  .5
                                        s4_a  .4  a2_e  .5
                                                  d2_c  .5
                                                  d2_e  .5
                                                  e1_c  .5
                                                  s1_d  .5
                                                  s2_a  .5
                                                  s4_e  .5
    .##########  |            c1_a  .3  c2_c  .4  a1_a  .5
                                        s3_d  .4  a1_d  .5
                                        s4_c  .4  a2_f  .5
                                                  d1_a  .5
                                                  d2_a  .5
                                                  e2_a  .5
                                                  s4_f  .5
  .############ M|                      c1_c  .4  b1_c  .5
                                        c1_d  .4  b2_b  .5
                                        c2_d  .4  d1_c  .5
                                        d4_b  .4  e1_b  .5
                                        s3_a  .4  e2_c  .5
                                                  s1_c  .5
   .###########  |S                     b1_e  .4  b1_b  .5
                                        d1_b  .4  d2_b  .5
                                        d1_f  .4  e1_d  .5
                                        d4_c  .4  s2_d  .5
                                        s2_e  .4
  .############  |  c1_a  .2  c1_b  .3  a2_c  .4  b2_a  .5
                    c2_a  .2  c2_b  .3  b1_d  .4  d2_d  .5
                              s1_e  .3  b2_d  .4  e2_d  .5
                              s4_a  .3  d2_f  .4  s1_a  .5
                                        d3_b  .4
                                        d3_c  .4
                                        d3_f  .4
                                        d4_a  .4
                                        d4_d  .4
                                        d4_e  .4
                                        d4_f  .4
                                        s3_b  .4
                                        s4_b  .4
     .#########  |            c2_c  .3  a2_a  .4  a2_g  .5
                              s3_d  .3  a2_b  .4  e2_b  .5
                              s4_c  .3  b2_c  .4  s2_c  .5
                                        d1_d  .4
                                        d1_e  .4
                                        d3_d  .4
                                        d3_e  .4
                                        e1_a  .4
                                        s3_c  .4
     .#########  |            c1_c  .3  a1_c  .4  b1_a  .5
                              c1_d  .3  a2_d  .4
                              c2_d  .3  a2_e  .4
                              s3_a  .3  d2_c  .4
                                        d3_a  .4
                                        s1_b  .4
                                        s4_d  .4
                                        s4_e  .4
0     .########  +M c1_b  .2  b1_e  .3  a1_b  .4  a1_g  .5
                    c2_b  .2  d1_f  .3  a1_e  .4
                              d4_b  .3  a1_f  .4
                              s2_e  .3  a2_f  .4
                                        d2_e  .4
                                        e1_c  .4
                                        e2_a  .4
                                        s1_d  .4
                                        s2_a  .4
        .###### S|  c2_c  .2  a2_c  .3  a1_a  .4
                    s1_e  .2  b1_d  .3  a1_d  .4
                    s4_a  .2  b2_d  .3  d1_a  .4
                              d1_b  .3  d1_c  .4
                              d2_f  .3  d2_a  .4
                              d3_b  .3  e1_b  .4
                              d3_f  .3  s4_f  .4
                              d4_c  .3
                              d4_d  .3
                              d4_e  .3
                              d4_f  .3
                              s3_b  .3
                              s4_b  .3
          .####  |  c1_c  .2  a2_a  .3  b1_c  .4  s2_b  .5
                    s3_d  .2  a2_b  .3  b2_b  .4
                    s4_c  .2  b2_c  .3  e1_d  .4
                              d1_d  .3  e2_c  .4
                              d1_e  .3  s1_c  .4
                              d3_c  .3  s2_d  .4
                              d3_d  .3
                              d3_e  .3
                              d4_a  .3
                              e1_a  .3
```

```
                               s3_c  .3
             .##  |  b1_e  .2  a1_c  .3  b1_b  .4
                     c1_d  .2  a2_d  .3  d2_b  .4
                     c2_d  .2  d2_c  .3  d2_d  .4
                     d1_f  .2  d3_a  .3
                     d4_b  .2  s1_b  .3
                     s2_e  .2  s4_d  .3
                     s3_a  .2
              .#  |S b2_d  .2  a1_b  .3  a2_g  .4
                     d1_b  .2  a1_e  .3  b2_a  .4
                     d2_f  .2  a1_f  .3  e2_b  .4
                     d3_b  .2  a2_e  .3  e2_d  .4
                     d3_f  .2  a2_f  .3  s1_a  .4
                     d4_c  .2  d2_e  .3
                     d4_d  .2  e1_c  .3
                     s3_b  .2  e2_a  .3
                     s4_b  .2  s1_d  .3
                               s2_a  .3
                               s4_e  .3
              .#  |  a2_b  .2  a1_a  .3  s2_c  .4
                     a2_c  .2  a1_d  .3
                     b1_d  .2  d1_a  .3
                     d1_d  .2  d2_a  .3
                     d3_c  .2  s4_f  .3
                     d3_e  .2
                     d4_a  .2
                     d4_e  .2
                     d4_f  .2
                     s3_c  .2
               .  |  a2_a  .2  b1_c  .3  a1_g  .4
                     a2_d  .2  b2_b  .3  b1_a  .4
                     b2_c  .2  d1_c  .3
                     d1_e  .2  e1_b  .3
                     d3_a  .2  e1_d  .3
                     d3_d  .2  e2_c  .3
                     e1_a  .2  s1_c  .3
                     s4_d  .2  s2_d  .3
-1             . T+T a1_c  .2  b1_b  .3
                     a1_e  .2  d2_b  .3
                     a1_f  .2  d2_d  .3
                     a2_e  .2
                     d2_c  .2
                     d2_e  .2
                     s1_b  .2
                     s1_d  .2
                     s4_e  .2
               .  |  a1_a  .2  a2_g  .3  s2_b  .4
                     a1_b  .2  b2_a  .3
                     a1_d  .2  e2_b  .3
                     a2_f  .2  e2_d  .3
                     d2_a  .2  s1_a  .3
                     e1_c  .2
                     e2_a  .2
                     s2_a  .2
               .  |  b2_b  .2  s2_c  .3
                     d1_a  .2
```

```
                       d1_c  .2
                       e1_b  .2
                       e2_c  .2
                       s4_f  .2
                 .  |  b1_c  .2  b1_a  .3
                       e1_d  .2
                       s1_c  .2
                       s2_d  .2
                 .  |  b1_b  .2  a1_g  .3
                       d2_b  .2
                       d2_d  .2
                       e2_d  .2
                       s1_a  .2
                 .  |  a2_g  .2  s2_b  .3
                       b2_a  .2
                       e2_b  .2
                       s2_c  .2
                 .  |
  -2             .  +  a1_g  .2
                       b1_a  .2
                    |
                 .  |  s2_b  .2
                 .  |
                 .  |
                    |
                 .  |
  -3                +
              <less>|<freq>
EACH "#" IS 14: EACH "." IS 1 TO 13
```

Appendix C: Item Difficulty Measures

Table C.1. Item statistics: Measure order

Item Label	Score	Count	Measure	Model S.E.	Estimated Discrimination
c2_a	590	452	1.385	0.05	0.58
c1_a	593	436	1.326	0.05	0.57
c2_b	807	453	0.890	0.05	0.89
c1_b	804	437	0.833	0.05	0.71
s4_a	3406	1843	0.826	0.02	0.56
s1_e	3558	1877	0.771	0.02	0.93
c2_c	890	453	0.705	0.05	0.98
s4_c	3719	1844	0.656	0.02	0.84
s3_d	3761	1851	0.641	0.02	1.07
c1_c	904	435	0.604	0.05	0.87
c2_d	972	454	0.531	0.05	0.90
c1_d	942	434	0.516	0.05	0.78
s3_a	4012	1861	0.515	0.02	1.22
d4_b	978	443	0.465	0.05	1.20
b1_e	1026	464	0.459	0.05	0.60
s2_e	4175	1863	0.430	0.02	0.93
d1_f	1043	462	0.410	0.05	0.75
d1_b	1081	463	0.336	0.05	1.02
d4_c	1040	443	0.323	0.05	1.29
d3_b	1019	431	0.312	0.05	0.96
d3_f	1032	432	0.294	0.05	0.70
s4_b	4427	1848	0.274	0.02	1.08
d4_d	1061	443	0.273	0.05	1.12
s3_b	4462	1860	0.268	0.02	1.26
d2_f	1110	460	0.266	0.05	0.74
b2_d	1120	465	0.264	0.05	0.97
a2_c	1111	456	0.226	0.05	1.21
d4_e	1081	442	0.220	0.05	1.33
b1_d	1137	462	0.208	0.05	0.90
d4_f	1093	443	0.196	0.05	0.91
d4_a	1094	443	0.192	0.05	1.07

Item Label	Score	Count	Measure	Model S.E.	Estimated Discrimination
d3_c	1077	433	0.184	0.05	1.21
s3_c	4628	1859	0.174	0.02	1.33
a2_b	1138	457	0.170	0.05	1.12
d1_d	1150	461	0.167	0.05	1.11
d3_e	1099	435	0.146	0.05	1.24
d1_e	1186	463	0.097	0.05	1.43
b2_c	1203	467	0.091	0.05	0.93
e1_a	1173	456	0.084	0.05	0.70
a2_a	1176	457	0.084	0.05	0.97
d3_d	1120	431	0.074	0.05	1.19
s4_d	4843	1848	0.034	0.02	1.10
a2_d	1196	454	0.012	0.05	1.22
d3_a	1150	434	0.010	0.05	0.95
a1_c	1174	437	-0.036	0.05	1.21
s1_b	5072	1879	-0.060	0.03	1.09
d2_c	1260	462	-0.080	0.05	1.15
s4_e	5056	1850	-0.092	0.03	1.24
a2_e	1245	455	-0.102	0.05	1.21
d2_e	1284	464	-0.125	0.05	1.33
s1_d	5203	1877	-0.143	0.03	1.16
a1_e	1214	436	-0.146	0.05	1.23
a1_f	1226	439	-0.161	0.05	0.90
s2_a	5250	1874	-0.180	0.03	0.83
a1_b	1236	439	-0.188	0.05	0.76
e1_c	1296	460	-0.196	0.05	1.05
e2_a	1316	463	-0.215	0.05	0.27
a2_f	1296	457	-0.220	0.05	0.92
a1_d	1260	438	-0.260	0.05	0.99
a1_a	1272	440	-0.279	0.05	0.97
d2_a	1344	463	-0.288	0.05	1.07
d1_a	1352	463	-0.324	0.05	1.04
s4_f	5421	1850	-0.335	0.03	0.92
d1_c	1366	461	-0.381	0.05	1.23
e1_b	1364	460	-0.386	0.05	0.94
b2_b	1390	465	-0.406	0.05	0.89
e2_c	1409	468	-0.434	0.05	0.85
s1_c	5682	1877	-0.469	0.03	0.75
b1_c	1407	464	-0.472	0.06	0.98
e1_d	1399	459	-0.496	0.06	0.79
s2_d	5713	1872	-0.503	0.03	1.17
b1_b	1448	464	-0.599	0.06	0.79

Item Label	Score	Count	Measure	Model S.E.	Estimated Discrimination
d2_b	1459	464	-0.623	0.06	1.00
d2_d	1464	463	-0.653	0.06	0.97
e2_d	1496	468	-0.708	0.06	0.79
s1_a	5998	1877	-0.718	0.03	1.13
b2_a	1502	465	-0.772	0.06	1.05
e2_b	1518	468	-0.786	0.06	0.80
a2_g	1469	452	-0.803	0.06	0.56
s2_c	6123	1876	-0.828	0.03	0.86
b1_a	1570	465	-1.031	0.07	0.87
a1_g	1494	438	-1.099	0.07	0.78
s2_b	6594	1873	-1.347	0.04	1.04

Appendix D: Additional Scenarios

In reading the prompts below, place yourself in the scenario:

You are in the common area of your residence hall talking with some friends when you observe someone from your floor, who you don't know very well, writing something on a door of a friend. As this person starts to walk away, you notice that they wrote a racial slur directed at your friend.

How likely are you to:

a. Say something at the time to the student who wrote the comment?
b. Say something at the time to your friend who was targeted?
c. Say something at a later time to the student who wrote comment?
d. Say something at a later time to your friend who was targeted?
e. Get other people to support you in intervening?
f. Get an authority figure (RA, police, residence hall/apartment manager, etc.) to intervene?
g. Remove what they wrote?

In reading the prompts below, place yourself in the scenario:

You are enrolled in a course you really like with a great professor you haven't had before. One day during lecture, your professor makes a few jokes based on religious stereotypes. You don't have any other friends in the class.

How likely are you to:

a. Say something at the time to your professor in front of the class?
b. Say something at a later time to your professor in private?
c. Get other people to support you in intervening?
d. Say something at a later time to an authority figure (another professor, campus staff, etc.)?

In reading the prompts below, place yourself in the scenario:

You are on the executive board of your student organization. During a recent meeting, one of the members makes a comment that everyone should invite only opposite-gender partners to the upcoming social.

How likely are you to:

a. Say something at the time to the member who made the comment?
b. Say something at a later time to the member who made the comment?
c. Get other people to support you in intervening?
d. Get another leader of your student organization to intervene?

In reading the prompts below, place yourself in the scenario:

You are in the dining hall eating with some friends when a group of students you don't know sits at a nearby table. When they begin to talk to each other in a language other than English, one of your friends makes a comment that this is America and everyone should speak English.

How likely are you to:

a. Say something at the time to your friend who made the comment?
b. Say something at a later time to your friend who made the comment?
c. Get other people to support you in intervening?

In reading the prompts below, place yourself in the scenario:

You are riding a crowded campus bus to your next class when a student using a wheelchair gets on. It takes them a while to get in a safe position for riding, and people on the bus start to complain.

How likely are you to:

a. Say something at the time to the people making comments?
b. Say something at a later time to the student in the wheelchair?
c. Get other people to support you in intervening?
d. Get the bus driver to intervene?

In reading the prompts below, place yourself in the scenario:

You are with a friend in the locker room of the campus recreation center when you overhear someone tell another person that they are in the wrong room based on their sex and should go to the other locker room.

How likely are you to:

a. Say something at the time to the person who made the comment?
b. Say something at the time to the person who was asked to leave?
c. Get other people to support you in intervening?
d. Get someone from the rec center staff to intervene?