ABSTRACT

Title of Dissertation:	THE CRITICAL RACE FRAMEWORK STUDY: STANDARDIZING CRITICAL EVALUATION FOR RESEARCH STUDIES THAT USE RACIAL TAXONOMY
	Christopher Menvell Williams, Doctor of Philosophy, 2024
Dissertation directed by:	Dr. Craig S. Fryer, Associate Professor, Department of Behavioral and Community Health

Race is one of the most common variables in public health surveillance and research. Yet, studies involving racial measures show poor conceptual clarity and inconsistent operational definitions. There does not exist a bias tool in the public health literature for structured qualitative evaluation in critical areas of critical appraisal – reliability, validity, internal validity, and external validity – for studies that use racial taxonomy. This study developed the Critical Race (CR) Framework to address a major gap in the literature.

The study involved three iterative phases to answer five research questions (RQs). Phase I was a pilot study of the CR Framework among public health faculty and doctoral students to assess measures of fit (RQ1) and to identify areas of improvement in training, instrumentation, and study design (RQ2). Study participants received training and performed a single article evaluation. Phase II was a national cross-sectional study of public health experts to assess perceptions of the revised training and tool to assess measures of fit (RQ1), to determine the influence of demographic and research factors on perceptions (RQ3), and to gather validity

evidence on constructs (RQ4). In Phase III, three raters performed article evaluations to support reliability evidence (RQ4) and to determine the quality of health disparities and behavioral health research studies against the CR Framework (RQ5).

We assessed the reliability of study results and the CR Framework using non-differentiation analysis, thematic analysis, missingness analysis, user data, measures of internal consistency for adopted instruments, interrater agreement, and interrater reliability. Validity was assessed using content validity (CVI and *k**), construct validity, and exploratory factor analyses (EFA).

The study recruited 30 highly skilled public health experts across its three phases as part of the final analytic sample. Phase I had poor reliability in which the results could not be confidently interpreted (RQ1) and indicated needed improvement in study design, training, and instrumentation (RQ2). Based on Phase II results, we met or exceeded acceptable thresholds for measures of fit – acceptability, appropriateness, feasibility, and satisfaction (RQ1). Demographic or research factors were not associated with responses (RQ3). Interrater agreement was moderate to high among rater pairs (RQ4). Due to lack of confidence in significance testing, interrater reliability results were inconclusive. Overall data results showed excellent content validity. Based on EFA results, construct validity for reliability and validity items was poor to fair (RQ4). Data results were inconclusive on internal validity and external validity. The twenty studies used in critical appraisal showed low quality or no discussion when the Critical Race Framework was used (RQ5).

The CR Framework study developed a tool and training with quality evidence for implementation effectiveness, content validity, and interrater reliability to fill a major gap in the public health literature. It contributed an innovative theory-based tool and training to the

literature. Future research should seek to study individual perceptions and practices that influence outcomes of CR Framework application and to reduce barriers to ensure that minimum sample sizes can be met for additional testing.

THE CRITICAL RACE FRAMEWORK STUDY: STANDARDIZING CRITICAL EVALUATION FOR RESEARCH STUDIES THAT USE RACIAL TAXONOMY

by

Christopher Menvell Williams

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Associate Professor, Dr. Craig S. Fryer, Chair
Associate Professor, Dr. James Butler III
Assistant Research Professor, Dr. Devlon N. Jackson
Professor, Dr. Mia Smith-Bynum, Dean's Representative
Professor, Dr. Min Qi Wang

Dedication

I dedicate this dissertation to the women leaders in my community of practice – Mrs. P. Bishop, Mrs. L. Brown, Commissioner R. Hamilton, Mrs. C. Spencer, and Mrs. D. Walker. Your community leadership and lifelong service in Washington, DC enriched my understanding of contemporary structural racism and health inequity reproduction. You provided an essential part of my professional identity and training. Together, we published our manuscript on elucidating and affecting the *public health economy* (C. Williams, Birungi, et al., 2022). The policies, poor agency performance, and societal attitudes that perpetuate particularized harm against vulnerable communities deserve heightened scrutiny in public health research, policy, and practice. The ostensible concordance with those to whom we targeted our advocacy and research compelled me, as a researcher, to directly confront major drawbacks in public health theory and practices. The current paradigm not only impedes scientific progress, but also obscures unique community experiences like yours that explain the persistence of vast health inequity and structural racism. Families in low-income housing warrant sustained focus in the discipline of public health.

Acknowledgements

I wish to acknowledge and give my deepest appreciation to my aunt, F. Williams, Ph.D. Her mentorship throughout my doctoral training and dissertation served an invaluable role in making this academic achievement possible. My parents, R. Love and S. Williams, instilled in me strong moral conviction and discerning judgment that propel my research and advocacy. To my friends, thank you for social and emotional support during this incredible journey. My friendships and community leaders mean so much.

Throughout my educational journey, I have had close personal relationships with many teachers who provided immense support and encouragement. I owe them a debt of gratitude. I especially want to thank Ms. J. Campbell, Mrs. C. Currence, Ms. S. Wright, Professor M. Smith, and Mrs. King. Dr. J. El-Bayoumi, Dr. S. Angus, Dr. E. Muchmore, and Dr. R. Alweis opened new opportunities in research training that inspired my educational pursuits. I am deeply grateful to my supervisors: Mrs. J. Hilton, Mrs. D. Farineau, Dr. J. Donowitz, and Mrs. R. Overton.

I am grateful to my academic advisor and dissertation committee chair, Associate Professor Craig Fryer, Dr.P.H., who navigated me through five years of doctoral studies. A debt of gratitude is owed to my dissertation committee for supporting this important area of research inquiry.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CR. Critical Race

EFA. Exploratory Factor Analysis

ICC. Intraclass Correlation Coefficient

IRR. Interrater reliability

OMB. Office of Management and Budget

PHCRP. Public Health Critical Race Praxis

UMD. University of Maryland

Chapter 1: Background, Context, and Theoretical Framework

1. Introduction to Chapter 1

Race variables appear frequently in public health research (T. A. LaVeist, 2005; Ross et al., 2020). Yet, the scientific basis for use of race in health research has been a long-standing contentious issue (Gannon, 2016; Ioannidis et al., 2021; Lin & Kelsey, 2000; Witzig, 1996) Its common use is attributed to research norms rather than scientific rigor (Lin & Kelsey, 2000; Witzig, 1996). There is no general agreement on the meaning and relevance of race for public health science (Witzig, 1996). Studies that use a race variable show poor conceptual clarity and limited discussion of racial constructs consistent with scientific rigor (Martinez et al., 2022). Systematic approaches to evaluate this inherent threat to research quality are limited. A single bias tool developed in the medical education field appears in the literature (Garvey et al., 2022). A more structured tool to include conceptualization, analysis, and interpretation is needed to assist public health researchers to identify weaknesses in health studies. No such tool appears in the public health literature. The Critical Race (CR) Framework is a vital step to advance evidence-based reasoning and critical appraisal for health disparities research. This dissertation study developed a web-based training and tool to aid public health experts in evaluating health studies in four inviolable areas of critical appraisal – reliability, validity, internal validity, and external validity. Chapter 1 discusses each of these areas in our primary theoretical framework. We briefly explain our research questions and appropriateness of our study design - survey and article critique. Our assumptions, delimitations, and limitations provide the scope of our study. This study addressed a major gap in the critical appraisal and health disparities literature.

2. Background

The scientific utility of race is experiencing increasing debate within the research community (Burchard et al., 2003; Feero et al., 2024; Garvey et al., 2022; D. S. Jones et al., 2024; Martinez et al., 2022; National Academies of Sciences, 2023). While Critical Race theorists have sought to more centrally situate race in health research, critics have questioned the importance of racial data altogether (Butler et al., 2018; Ford & Airhihenbuwa, 2018; Gannon, 2016; Garvey et al., 2022; Ioannidis et al., 2021; Witzig, 1996). Racial nomenclature forms a type of research bias toward favoring the legitimacy of race in science, except that researchers often fail to explain the construct of race that it is intending to capture (Kaufman & Cooper, 2001). This study sides with critics in arguing that race variables inherently weaken research quality and impedes scientific advancement (Fanelli & Ioannidis, 2013). Conceptual and operational definitions of race in research are highly attenuated, for reasons that we explain throughout this study.

The inclusion of racial variables is understood to benefit public health by aiding in determining causes of disease — risk factors, genetics, and the role of the environment (Bhopal & Donaldson, 1998; Lin & Kelsey, 2000). Support for race and ethnicity in public health research has been premised on four major assumptions: 1) that race and ethnicity are easily attainable and consistently defined, 2) that there exists shared interpretability among study populations and the general public, 3) that research participation and response rates are similar across groups, and 4) that race is a stable variable among research participants across data sets and periods of assessment (CDC, 1998). Some researchers believe that abandoning race would harm certain populations, “We believe that ignoring race and ethnic background would be detrimental to the very populations and persons that this approach allegedly seeks to protect.

Information about patients' ethnic or racial group is imperative for the identification, tracking, and investigation of the reasons for racial and ethnic differences in the prevalence and severity of disease and in responses to treatment" (Burchard et al., 2003). Sometimes, race science is used to support their position, as illustrated by this claim, "Despite the admixture, (B)lack Americans, as a group, are still genetically similar to Africans" (Burchard et al., 2003).

As discussed later, scientific reasoning to justify race must uphold standards that researchers apply to all other types of variables. Proponents of racial analysis often struggle to square with this rigor (Martinez et al., 2022). Race data collection tools should show reliability and validity. Threats to internal validity due to poor construction of race should be assessed and minimized to the greatest extent possible. External validity should support population, ecological, and temporal validity.

Critical Race Theory (CRT) strongly argues for intensifying the study of race. CRT is a school of thought that promotes critical analysis of race and racism across a variety of disciplines. It assumes that racism is endemic to US society and that white supremacy perpetuates systemic racism (Crenshaw, 2015; Ladson-Billings & Tate, 1995; Stovall, 2005; West, 1995). It has its origins within Critical Legal Studies wherein purported color blindness in the law was challenged (Crenshaw, 2015; Delgado & Stefancic, 2001). CRT has been adopted in many disciplines. The mid-nineties marked its introduction in education, followed by public health research in 2010 as "Principles of Public Health Critical Race Praxis (PHCRP) (Ford et al., 2010; Ladson-Billings & Tate, 1995). CRT posits a conceptual and methodological orientation that upholds the centrality of race (Ford et al., 2010).

Although the CRT literature is highly context-dependent and not a theory in the traditional sense of a concretized theory of change, there are at least four generally accepted

principles: race consciousness, contemporary mechanisms, centering in the margins, and praxis (Ford et al., 2010). Race consciousness seeks to centrally and contextually place race and racism discourse – whether historical, legal, educational, public health, and research (Ford et al., 2010). Contemporary mechanisms draw attention to the ways in which race and racism impact decisions and policies within these fields, heavily relying on the social context (Ford & Airhihenbuwa, 2018). Centering brings marginalized perspectives and lived experiences to the forefront of discourse and research (Ford et al., 2010). Praxis concerns the synergism of the prior principles to transform public health practice (Ford et al., 2010).

When Ford and colleagues introduced CRT in public health research, PHCRP built upon the foundational principles of CRT: public health, centering, critical consciousness, experiential knowledge, ordinariness, praxis, primacy, race consciousness, and the social construction of minoritized populations (Ford et al., 2010). The first PHCRP-based training targeting academic researchers was published in 2018 and showed promise as a training model (Butler et al., 2018). The term Quantitative Critical Race ("QuantCrit") theory is often used when CRT is applied in statistical data analysis and interpretation (Castillo & Gillborn, 2022).

After more than twenty years of CRT, "this promise of a coherent account of race had not yet come to fruition" (Cabrera, 2019). Although its supporters understand CRT as an intellectual movement, it has attracted much criticism. Critics have argued against CRT on several fronts: does not contain the "intellectual architecture" to be considered social theory, lack of clear mechanistic racial theory, overemphasis and underdevelopment of systemic racism and white supremacy as root causes, lack of methodological translation, allows too much intellectual variety, lack of Latinx inclusion, and non-agreement among CRT adherents (Anguiano & Castañeda, 2014; Cabrera, 2019; Treviño et al., 2008). Based on a PubMed search conducted on

November 2, 2022, for the term "PHCRP," public health CRT remains in its formative stage, as little appears on standardized tools and trainings since the first published training conducted by Butler and colleagues (Butler et al., 2018).

Proponents for the use of race in research have often regarded race as a proxy for several constructs – income, culture, biology, and racism (C. P. Jones, 2000; T. A. LaVeist, 2005). In Jones and colleagues' highly cited work positing a theory on three levels of racism, they purported that race is a "rough proxy for socioeconomic status, culture, and genes, but it precisely captures the social classification of people in a race-conscious society such as the United States…That is, the variable "race" is not a biological construct that reflects innate differences, but a social construct that precisely captures the impacts of racism" (C. P. Jones, 2000). While Jones' position is that "race is a contextual variable, not a characteristic of the person," she acknowledges several issues with race while seeking to defend race as an indicator of "the distribution of risks and opportunities in our race-conscious society" (C. P. Jones, 2001). She presents race as without a shared cultural meaning ("There is no single Black culture, just as there is no single White, Hispanic, or Asian culture"), no genetic delineation ("There is no denying that there is genetic variability on the planet. However, the pie slicer that we call race does not capture that genetic variability"), and spatially and temporally changeable ("this assigned race may vary over time") (C. P. Jones, 2001). Despite these major limitations, she and her colleagues' research position can be understood as encouraging the centrality of race as a researcher's social responsibility to acknowledge and respond to inequity in a race-conscious social context.

Similar to Jones, Thomas LaVeist holds a complex view of race in public health research – acknowledging conceptual and methodological challenges. LaVeist's physiognomy model of

race and health of cause and effect posits three pathways for explaining racial health disparities – behavioral, cultural or ethnic, social or structural (T. A. LaVeist, 1994). He identified inherent weaknesses of race: poor proxy ("it seems logical that if race is a proxy for other factors such as biology or culture, then a need exists to find more creative ways to measure these other factors"), lack of practical translation ("Moreover, from a statistical standpoint the simple inclusion of a race dummy variable in a regression model is inadequate if the objective is to develop interventions to affect race differences in a dependent variable"), inadequate for population validity ("In practice, justified examples for using race as a criterion in sample selection are rare"), and measurement error ("there are measurement problems with race that have not been adequately addressed")(T. A. LaVeist, 1994). LaVeist concluded that, "Researchers should treat the race variable with the same degree of caution and skepticism with which it treats any other variable" (T. A. LaVeist, 1994) The latter is the underlying premise of this study. Still, LaVeist heavily relies on the construction of race in his body of work. For example, a study compared black and white residents in Baltimore, Maryland and found that neighborhood ("place") was misattributed to race (T. LaVeist et al., 2011). In other words, race is highly valuable for investigating health disparities despite considerable limitations.

On the other hand, CRT critics have posited that racial nomenclature lacks: 1) conceptual definitions, 2) a viable cultural proxy or surrogate, 3) clarity into disease etiology, 4) an adequate response to within-group heterogeneity, and 5) standardization (Bhopal & Donaldson, 1998; Fullilove, 1998; Ioannidis et al., 2021; Lin & Kelsey, 2000). A recent systematic review showed poor conceptual clarity of the race variable, lack of delineation between ethnicity and race, and limited discussion of how race was measured (Martinez et al., 2022).

The National Academies of Sciences, Engineering, and Medicine (NASEM) recently established a committee to evaluate the use of race and ethnicity in biomedical research (National Academies of Sciences, Engineering, and Medicine, n.d.). In 2023, NASEM released a report that has been interpreted as, “a tectonic shift away from current models that use race, ethnicity, and geographic origin as proxies for genetic ancestry groups (ie, a set of individuals who share more similar genetic ancestries) in genetic and genomic science” (Feero et al., 2024; National Academies of Sciences, 2023).

In response to NASEM, journal editors published ten percepts in March 2024 for authors and peer journal reviewers to consider when race and ethnicity are used as proxies for genetic ancestry groups that: 1) include accurate and respectful terminology and descriptors, 2) avoidance of race and ethnicity as genetic proxies, 3) expressed rationale for population groups, 4) methodologies for data collection, 5) operationalization of populations, 6) inclusion of all genetic ancestry groups, 7) how populations of interest influence data interpretability, 8) limitations of generalizability with stratified analysis, 9) limitations with use of legacy datasets, and 10) avoidance of race and ethnicity as proxies for genetic ancestry (Feero et al., 2024). Feero and colleagues did not address the use of race other than as proxies for genetics and genomics. It is therefore implied that race is less problematic if race research avoids tenuous linkages to genetic ancestry. In addition, their precepts appeared in the literature in March 2024 after the development and testing of our critical appraisal tool (“Critical Race Framework”). Their percepts are not presented as a critical appraisal tool.

Researchers have discouraged racial measures as a primary study measure, “It is questionable, therefore, whether this should ever be a quantity of interest for biomedical researchers, except in cases when race is a marker for a process external to individual

physiology, as in the investigation of health services disparities or other sociologic questions" (Kaufman & Cooper, 2001). There is a need for "developing a valid nomenclature for that (White) population and its subgroups; and demonstrating that data on subgroups of the White population can be successful in improving the health status of worse-off subgroups." (Bhopal & Donaldson, 1998). "I believe it is time to abandon race as a variable in public health research. Following the illusion of race cannot provide the information we need to resolve the health problems of populations" (Fullilove, 1998). Rather than a single trait or combination of traits, some researchers have called for narrative-based categorization, an anti-racist community-based participatory research (CBPR) model, and effectuation of the public health economy (Adkins-Jackson et al., 2023; Hsu et al., 2019; C. Williams, Birungi, et al., 2022).

Academic journals are culpable in widespread practices that assume data are representative and interpretative when they are not. An example is instructive. Xiao and colleges published a systematic review and meta-analysis of 122 COVID-10 clinical studies between October 2019 and February 2022 to investigate female and racial/ethnic minority enrollment in trials (Xiao et al., 2023). This study was published in the *Journal of the American Medical Association* (JAMA) – a highly prestigious medical science journal. Xiao and his colleagues concluded that, "female participants were underrepresented in treatment trials, Asian and Black participants were underrepresented in prevention trials, and Hispanic or Latino participants were overrepresented in treatment trials," except that these findings are hardly interpretable (Xiao et al., 2023). Xiao and colleagues' analytical framing of White, Black, Asian, Native Hawaiian, and American Indian "races" are emblematic of the common poor practices in public health research. Xiao and colleagues supported a racial lens in their study due to 1) racial differences in COVID-19 incidence, hospitalization, and mortality, 2) race as part of "independent modulators of

drug/vaccine efficacy and toxic effects in specific settings," and 3) historical minority underrepresentation in clinical trials, and 4) importance of representativeness of diverse populations (Xiao et al., 2023). They discuss many limitations in their study. The major sources of bias arising from racial social construction are not discussed and not accounted for in data analysis. They assume that the significance and homogeneity of racial categorizations are commonly held. Due to this study and much of health disparities research overlooking the scientific rigor necessary for race variables, the Critical Race Framework study is encouraging a major sea change in critical appraisal of research quality, including among top health and medical sciences journals.

Researchers have also encouraged greater focus on racial subgroups over homogeneous groups (CDC, 1998). This study does not take that position. The *supra*-construct of race implied in this framing remains a major problem for health disparities research, as discussed. It may be a position that is somewhat consistent with CRT because CRT encourages "deeper understandings of concepts, relationships, and personal biases" (Ford et al., 2010). However, the CRT assumption based on the primacy of race is a fundamental feature of its school of thought. In our study, we assume that race is an anachronistic hold-over that "developed largely to justify the highly profitable African slave trade and the systems of slavery in the Americas" (Fullilove, 1998). Our study premise is that the centuries-old social construction of race has devolved as be too attenuated and crude for public health research. It weakens research quality, encourages poor practices, and retards scientific progress.

3. Context

This dissertation study took place in the University of Maryland School (UMD) of Public Health Department of Behavioral and Community Health. All research activities were virtual –

two cross-sectional studies (Phases I, II) and article critiques among three researchers (Phase III).

4. Theoretical Framework

The primary theoretical framework derives from four inviolable principles for research quality and critical appraisal: reliability, validity, internal validity, and external validity. Critical appraisal is an approach and method for examining research to assess scientific quality, value, and practical significance (Al-Jundi & Sakka, 2017). It is vital for researchers to possess a comprehensive understanding of threats in these four areas in conducting their own research and evaluating the quality of studies. Poor quality research wastes resources, retards scientific progress, and poses threats to public health and safety (Fanelli & Ioannidis, 2013).

5. Problem Statement

Studies involving racial analysis show poor conceptual clarity, non-delineation between ethnicity and race, and minimal discussion of quality of racial data collection tools (Martinez et al., 2022). It was not known the extent to which the development of a bias tool and training would be acceptable, feasible, and appropriate. We did not know whether research and teaching experience, demographics, or perceptions of our study premise would influence study results. It was also not known how highly cited disparities studies would rate against such a tool to qualify study bias due to the inclusion of race conceptualization and analysis. A similar tool that was recently published in the medical education literature only involved a single article critique (Garvey et al., 2022). The Critical Race Framework addresses a major gap in the critical appraisal literature in public health. Health disparities research informs public health research and practice in educational, policy, and regulatory arenas. The significance of this study is apparent and compelling. Without critical reflection of race in research, the highest quality

research and evidence-based practices cannot be achieved. The original tool and training arising from this study are likely to contribute to scientific advancement.

This study is situated within broader critiques and critical positionality to research norms and practices. Critics have decried misinterpretation, underreporting, and abuse in statistical testing and research (Greenland et al., 2016; Lang et al., 1998). Much criticism has been directed at whether model assumptions of each statistical method are adequately tested and satisfied prior to data analysis. Researchers commonly mistake *p*-values for importance and credibility (Greenland et al., 2016; Lang et al., 1998). In other words, research requires both quantitative and qualitative judgment. In Chapter 2, we discuss statistical tests that may or may not be appropriate depending on the theoretical framework for race or racial health disparities. Differences in social or structural factors are commonly ascribed to health disparities. In such instances, data analysis should match those theories. We later elaborate on alternate frameworks and analyses. The underestimated and attenuating effects of race on research quality and interpretability are equally deserving of critical appraisal.

6. Purpose of the Study

The purpose of this study was to develop an acceptable, appropriate, relevant, and feasible bias tool to structure qualitative critical appraisal and that provided evidence of reliability and validity. Our research methodology for Phase I involved a survey and single article critique among public health faculty and doctoral students at a single site to study *measures of fit* – acceptability, feasibility, appropriateness, and relevance – related to the Critical Race Framework. A second major aim of Phase I was to capture quantitative and qualitive data to refine the study design, instrumentation, and training for Phase II. The Phase II research methodology was a national cross-sectional survey among public health experts to gather data on

measures of fit and validity evidence of the CR Framework. The research methodology for Phase III was a series of article critiques using the CR Framework among three researchers to gather reliability evidence.

7. Research Questions

This study defined five research questions (RQs) that are directly aligned with the problem statement. First, a bias tool was needed in public health that fully accounted for four critical areas of critical appraisal – reliability, validity, internal validity, and external validity. The only similar tool in the literature had many limitations and generally supported current practices. We sought to ensure development of a tool grounded in theory and supported by measures of implementability – our *measures of fit*. Our research questions first gathered data on a preliminary framework (RQ1) for which we identified areas of improvement in study design, instrumentation, and training (RQ2). Consistent with pilot studies, RQ1 sought to test the feasibility of implementing an innovation and the likelihood of challenges with subsequent research stages (In, 2017)**.** We then sought reliability (RQ4) and validity (RQ5) evidence against which to measure the quality of our critical appraisal tool. We also needed to assess potential factors associated with survey measures and perceptions (RQ3).

RQ1. To what extent is the CR Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to *a priori* constructs.

RQ2. What are needed improvements to the CR Framework 1.0 study design, instrument, and training?

RQ3. What are demographic and research factors associated with perceptions of the CR Framework?

RQ4. What is the reliability and validity evidence for the CR Framework?

RQ5. What is the quality of 1) highly cited health disparities studies and 2) behavioral health studies based on the CR Framework?

8. Rationale for the Study

The rationale for this study is that innovation in critical analysis is needed to qualify the threats to public health research due to practices in racial nomenclature, data collection, analysis, and interpretation. The development of a web-based training and structured approach using the Critical Race Framework is an important step to advance this field of study. Our study methodologies are the best approach to answer our research questions because experts are essential to determining whether the CR Framework is theoretically linked, implementable, acceptable, and appropriate (Proctor et al., 2011; Weiner et al., 2017). We worked within real-world constraints due to a limited budget and time constraints. The phasing of our study was needed to account for realistic resources and expectations for our participants to complete a set of prescribed research activities accurately and timely.

9. Significance to Public Health

This dissertation study has several implications for quality research, evidence-based policy, and public health practice.

Research. Quality health research is essential for evidence-based policymaking, including public health spending. The approval of drugs, vaccines, medical devices, and diagnostic standards relies on quality research that accounts for bias to determine efficacy and appropriate treatment. Variables used in statistical testing need to demonstrate reliability and validity. Race is the second most used variables in health research (Ross et al., 2020). Yet, there is no general agreement on its meaning and relevance (Witzig, 1996). This creates several

challenges for establishing the best quality research. Improving critical analysis is essential to high quality research.

The implications of this study mean that researchers should seek quality representations. Although it is the beyond the scope of the dissertation stage of the CR Framework study, there is an urgent need to identify significant associations and characteristics that bind people together and to be explicit in using that framework throughout data collection and analysis. These groups may be historical, sociocultural, religious, doctrinal, regional, or ideological taxonomies – or a combination of these.

In addition, there is a more practical research implication. High redundancy has been shown in biomedical research (Lund et al., 2022). Disaggregation of racial data can lead to place- and attribute-based populations that vary in severity of health needs and disparities. The author (CW) has illustrated that public housing communities can suffer from immense and unique public health challenges – environmental racism, deteriorating housing quality, displacement, poor air quality, and abject failures in government agency performance (C. Williams, Birungi, et al., 2022). A shift from racial homogeneity can spur research innovation to study and improve the health of highly vulnerable populations that shoulder the burden of inequalities in the public health economy.

Policy. The successful development of a CR Framework would likely lead to new funding expectations on racial data collection and analysis to ensure quality research. Following full post-study testing, we anticipate that new policies and standards in data collection, grant funding, and scientific advancement will account for a paradigmatic shift in critical race analysis.

Public Health Practice. The implication of this study is that practices dependent on racial generalizations constitute poor practice. As discussed further in Chapter 2, the assumption of racial homogeneity is all too common despite the breadth of diversity within and across races.

10. Significance to Behavioral and Community Health Science

Research – Health behavioral research receives considerable public and private funding. Research that relies on faulty assumptions wastes resources (Fanelli & Ioannidis, 2013). Studies that make conclusions based on racial generalizations in behavioral health have limited scientific value and generalizability (Kaufman & Cooper, 2001). Further, research that does not account for the inherent threat of racial variables is likely to have lower quality.

Practice – Health behavioral program planning and implementation that rely on racialized science lacks appropriateness. Race is a social construct rooted in justifications for slavery and white supremacy ideology (Eigen & Larrimore, 2012). Race is highly attenuated, lacking clarity for public health practice. By extension, it suggests that public health practice perpetuates an inappropriate and potentially offensive framing, absent a clear and meaningful construction of race.

Policy – Government policy changes to support behavioral health, whether in funding, guidelines, or regulations, are expected to be based on sound science and evidence. Federal standards on race data collection warrant intensified and sustained scrutiny. Policies may be unjustified where racialized research undermines scientific quality.

11. Nature of the Study

This study had two survey designs – survey and article critique – to gather data among public health experts for assessing *measures of fit* (e.g., acceptability), reliability, and validity. Phase I was a pilot study of the Critical Race Framework 1.0, followed by a national cross-

sectional survey of the Critical Race Framework 2.0. Phase III involved three public health researchers applying the Critical Race Framework 3.0 to twenty public health studies. We collected data on self-rated familiarity with key research concepts, research experience, demographics, and attitudes to determine associations with perceptions of the CR Framework. Consistent with the literature, experts were consistently engaged in the development of our critical appraisal tool (Downes et al., 2016).

12. Definition of Terms

The following terms are used conceptually and operationally in this study (Table 1).

Table 1: Definition of Terms

Term	Definition
Reliability	The accuracy of the data collection tool to provide consistent results and can be replicated
Validity	The adequacy of the tool to capture a related construct
Internal validity	The ability to make causal inferences
External validity	The extent to which the study results are generalizable
Measurement error	Deviation from a true value
Interrater agreement	Agreement or consensus among raters
Interrater reliability	The reproducibility or consistency among raters
Appropriateness	Refer to relevance, fit, and compatibility of the CR Framework
Feasibility	Extent to which the CR Framework can be successfully used
Acceptability	Extent to which framework is agreeable, palatable, or satisfactory

13. Assumptions, Limitations, and Delimitations

The following assumptions were present in this study:

1. It was assumed that racial variables inherently introduce threats to research quality.
2. It was assumed that published research involving race is flawed due to research norms and practices that lack major consideration of *Assumption 1*.
3. It was assumed that this study could attract an adequate number of public health researchers to perform research activities and accept the premise of the study.
4. It was assumed that researchers could separate personal opinions or positions on race to apply scientific reasoning in the context of this study.
5. It was assumed that race does not derive from scientific reasoning.
6. It was assumed that Assumptions 1-5 did not constitute researcher or study bias.

The following limitations/delimitations were present in this study:

1. Each construct covered by the Critical Race Framework occupies a robust field of study that the Critical Race Framework did not intend to fully capture in the bias tool.
2. A delimitation of the CR Framework was that it was not intended for novice researchers.
3. This study was limited by practical challenges given resource limitations, time intensity of study activities, and limited funding.
4. A delimitation of this study was that it focused on whether studies that use racial taxonomy, race data collection and analysis conform to high scientific standards and meaningfully contribute to public health research and innovation.
5. A delimitation was that we applied standards for scientific rigor to health studies that use race. These are standards that we would apply to any variable in critical appraisal, only that our study drew attention to the threat inherency of race in research quality. Other

disciplinary theories or approaches about race in research are only relevant to subsequent stages of the Critical Race Framework study if they can contribute to reliability or validity and successfully defend internal and external validity in current practices. We limited our quantitative and qualitative judgment to public health research practices in data collection, analysis, and interpretation.

6. A delimitation of this study concerned the practical challenges (e.g., the number of article critiques) that reduced the extent of reliability and validity analyses to support the Critical Race Framework.
7. This study had delimitations because of limited literature on a conceptual and methodological framework for a single principal investigator to direct development of our critical appraisal tool.
8. A delimitation of this study is that it was not intended to explore what human categorizations would provide scientific utility for public health research.
9. A delimitation of this study was that the Critical Race Framework did not aim to quantify bias estimates pertaining to error or uncertainty in race analysis and interpretation.

14. Summary and Organization of the Remainder of the Study

Chapter 1 discusses the scientific underpinnings for this study. Given the common use of race in health research, this dissertation is intended to address a major gap in the literature. Chapter 2 discusses critical theories and approaches on the use of race in health research and provides findings from several literature reviews, including behavioral health studies and seminal texts. Chapter 3 provides study methodology and procedures on each component of our three-phase study. Chapter 4 contains our study results by phase and research question. We conclude our study in Chapter 5, which outlines the strengths and weaknesses of the dissertation phase of the

Critical Race Framework Study. Our study found that twenty studies drawn from the health disparities and behavioral health literature showed low quality or no discussion when the Critical Race Framework was used.

Chapter 2: Literature Review

1. Introduction to Chapter 2

Race data collection practices in public health research derive from research norms rather than scientific rigor (Lin & Kelsey, 2000; Witzig, 1996). Many proponents of current practices claim that race is needed to understand the causes of disease and to target health resources (Bhopal & Donaldson, 1998; Lin & Kelsey, 2000). However, the purported social construction of race lacks an adequate operational definition in biomedical research. Without a valid theoretical framework in research, systematic error is likely to occur (Tabachnick et al., 2007). Health studies are defined by poor conceptual clarity of the race variable, lack of delineation between ethnicity and race, and limited discussion of how race was measured (Martinez et al., 2022). The literature review discusses positions in favor of and opposed to current practices on race data collection and analysis. We have posited that persistence of racial nomenclature in public health research is detrimental to the field.

That literature review on key concepts in research is followed by a discussion of our conceptual frameworks in developing the Critical Race Framework, both theoretically and operationally. A brief analysis of behavioral health studies highlights the general issues of poor racial conceptualization and analysis. Two systematic searches were conducted to determine whether any bias tool comparable to the Critical Race Framework existed and to identify bias tools used in critical appraisal. We conducted a search to identify seminal texts in internal validity and external validity, then assessed whether our study topic had been raised. We critically appraised the Behavioral Risk Factor Surveillance System (BRFSS) – one of the largest health surveys in the US since the 1980s. Finally, Chapter 2 addresses the gaps in the critical appraisal literature related to our study premise.

2. Theoretical Frameworks

In the context of race data collection, reliability pertains to the accuracy of the tool to provide consistent results while validity concerns the adequacy of the tool to capture a related construct on racial meaning (Heale & Twycross, 2015). Poor reliability and validity weaken internal validity. Strong internal and external validity are essential for determining the rigor, accuracy, and utility for health disparities studies (McDermott, 2011). In experimental studies, internal validity primarily pertains to the degree to which changes in the dependent variables result from the treatment; whereas, external validity is the extent to which the study is generalizable (Slack & Draugalis, 2001). The remaining discussion highlights critical areas of appraisal as related to our study premise on race.

I. Reliability and Race in Research

Measurement error is the difference between a measured observation and its true value. Measurement error can be shaped by systematic error (error measurement or assessment) or random error (naturally occurring error associated with chance)(Viswanathan, 2005). Random error affects reliability while systematic error effects validity (Barraza et al., 2019). Within the context of this study, random error means the chance that some racial data can be inaccurate. An example of systematic error might be a tool that forces a single racial identity for all study participants that causes underreporting of multiracial identities. The lack of a theoretical framework and inattention to the quality of race data collection and analysis can also cause systematic error.

Lack of measurement error is an assumption for multivariate statistics (Tabachnick et al., 2007). When independent variables have measurement error, they can cause attenuation bias toward zero or the null (Gokmen et al., 2022). Tabachnick and colleagues regard the lack of

measurement error, “a clear impossibility in most social and behavioral research” (Tabachnick et al., 2007). Measurement error affects the ability to reliably interpret study findings (Jin et al., 2013). “Virtually all measures (except those using relatively infallible indicators, such as measures of biological sex) include error components that reflect not only random factors but method variance (variance attributable to the method being used, such as self-report vs. interviews) and irrelevant but nonrandom variables that have been inadvertently included in the measure” (Westen & Rosenthal, 2003). Measurement error is a key concern in health research on the issue of race.

Multivariate statistics assume no measurement error. If there are inconsistencies in collecting race (e.g., different instruments, use of observer identification), an evaluation of reliability and subsequent adjustment for reliability errors may be necessary since error can increase the likelihood of Type I or Type II errors (Tabachnick et al., 2007). Data collection methods on defining participants’ race (e.g., self-identify, chart recall, electronic health record) within the same study also introduce errors in reliability errors.

Although racial identity is often considered to be constant throughout one’s lifetime, research findings challenge that assumption (Agadjanian & Lacy, 2021). Race reflects contingencies of culture, society, politics, migratory patterns, and time. The social construction of race is unstable — apt to social and political changeability. Hispanic or Latino survey respondents are commonly ambiguous on “race”. Forty-two percent selected “some other race” in the 2020 US census, an increase of 7.7 million from 2010 (U.S. Census Bureau, 2021). Thirty-percent of Hispanic adolescents in a national survey indicated “no race,” while 36% said “other” (Kao & Vaquera, 2006). This represents measurement error because the census does not adequately capture race for at least some respondents. Since at least 1993, there have been calls

for Hispanic to be listed as a category for race (CDC, 1998). During the Biden administration, the federal government began re-considering standards for data collection by adding new options for Middle Eastern/North African and Hispanic (Wang, 2022). During the Obama administration, these were first proposed, then abandoned during the Trump administration (Wang, 2024). The White House's Office of Management and Budget (OMB) published an announcement in the Federal Register on March 29, 2024 to its Statistical Policy Directive No. 15 to include these additional response options for "race and/or ethnicity (*Revisions to OMB's Statistical Policy Directive No. 15*, 2024).

Racial fluidity or switching is common (Ferguson, 2004; Waters, 2000). This occurs when individuals do not consistently self-report race. "The statistician does not consider a third option: that a member's true race is not the race she reports herself to be on any survey" (Root, 2008). Racial switching is estimated to range between 5% -12% within the US population (Agadjanian, 2022; Agadjanian & Lacy, 2021). Social position such as employment, household income, and friends can exert influence on racial identification (Jackson, 2012; Saperstein & Penner, 2012). Vargas (2015) found that income status contributed to changes in self-identification as white or Hispanic (Vargas, 2015). A switch in racial self-identification from nonwhite to white racial identity was found among some Republican voters in the 2016 presidential election (Agadjanian & Lacy, 2021). Ethnic and racial switching in the National Health and Nutrition Examination Survey has been estimated at forty-two percent of respondents (Hahn et al., 1996).

Data survey tools to collect race are highly inconsistent (Saperstein & Penner, 2012). Racial categories differ depending on the database or survey involved (Flanagin, Frey, Christiansen, & AMA Manual of Style Committee, 2021). When major shifts in self-defined race

occur with instrument changes such as adding an option for "other" or allowing multiple responses, they are likely an indication of poor reliability. Starting in 2000, the US Census began allowing selection of more than one racial identity. Self-reports of multiracial identity increased substantially (A. Brown, 2020). Individuals who reported multiple races increased from 2.9% to 10.2% between 2010 and 2020 (Bureau, n.d.). Excluding write-in responses, the multiple response option produced at least 62 possible race combinations – single "races" and 57 for two or more "races" (Parker et al., 2022).

II. Validity and Race in Research

Race and ethnicity are inherently ambiguous, subjective, and mutable (Cokley, 2007; Martinez et al., 2022). Research studies infrequently define race and ethnicity with any clear meaning (Martinez et al., 2022). Racial fluidity and switching suggest that racial meaning is likely too fluid, multi-faceted, and changeable (Tran & Johnston-Guerrero, 2016). The Centers for Disease Control and Prevention (CDC) described the mainstream support for race in health research, "We use the category of race as a means of identifying and measuring disparities. However, race has no biological meaning; it is a social construct designed intentionally to relegate people and communities of color to second-class status and to privilege white people" (CDC, 2023). This reasoning is not adequate for supporting current research practices. Arguably, it relies on circular logic: 1) disparities persist based on a "relegating" legacy of racial social construction, 2) thus, racialized research is justified, and 3) necessary for explaining disparities.

Race cannot meet the assumptions for construct validity: 1) that the variables are representative and exclusive, 2) reflect theory-based reasoning, 3) clearly understood before conducting the study, and 4) relies on a referent construct (Ferguson, 2004). The lack of validated tools impacts domains of construct validity: content, divergent, convergent, and

predictive (Goodwin & Goodwin, 1991). Further, Fowler identified features of construct measurement in implementation that included: 1) uniform and consistent interpretation, 2) consistent administration, including implementers' adherence to study procedures, 3) study participants access to information to make accurate responses, and 4) desire to provide an accurate answer at all times (Fowler, 1995). Translated in the context of this study, these standards of construct validity and measurement assume that: i) racial categories represent exclusive populations, ii) race represents a theory- and evidence-based construct prior to the start of a study, iii) race is consistently and uniformly interpreted by study respondents and the research team, iv) that study procedures provide information to participants about the study assumptions about race, and v) study participants are willing to accurately provide their race(s) based on the information available. These assumptions are not met in current research practices.

Race is an unstable construct historically. It is widely accepted that the establishment of the modern interpretation of race is a lingering legacy of the transatlantic slave trade (Forbes, 2009; Pierre, 2020). Race as a social construct did not begin to settle in common parlance until the 18th century when race-based slavery was codified (Eigen & Larrimore, 2012).

Biological Validity. Arguments in favor of a biological basis for race persisted well into the twentieth century and declined with the collapse of Nazi Germany (Bhopal & Donaldson, 1998). Scientific advancement in genetic testing appeared to settle the debate over biological race, but also coincided with its reemergence (Baker et al., 2017, 2017; Brattain, 2007; M. C. Campbell & Tishkoff, 2008; R. Frank, 2007). Genetic diversity is greater within races than among them, which deals a fatal blow to a biological basis for race (Freeman, 1998). Race does not represent a feasible proxy for genetic variation (Cooper, 2013; D. S. Jones et al., 2024).

Genetic research has perpetuated racial supremacy, particularly in the study of crime – a scientific practice that has been condemned (American Society of Human Genetics, 2018; Wedow et al., 2022). Racialized studies in criminology, to be sure, has been a long-standing issue in research, "However, over the past 30 years, the NIDA-funded focus on the prevalence and interrelationships of substance use and criminal involvement and the moderating influence of marriage may have also perpetuated some stereotypes, contributed to a continued focus in the research community on problem behaviors in Black communities, and unintentionally reinforced the importance of heteronormative institutions" (Doherty & Green, 2023). Survey research has shown mixed feelings among physicians for support for using race as in genomic medicine (Bonham & Nathan, 2002; Mateo & Williams, 2021).

US Census – The Official Language of Race. It is also commonly held that the US census is the official language of race (Hochschild & Powell, 2008). Table 2 lists the official races as

Table 2: Office of Management and Budget (OMB) Standards on Race – Five Races

- **American Indian or Alaska Native –** "A person having origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment."
- **Asian** – "A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam"
- **Black or African American** – "A person having origins in any of the Black racial groups of Africa."
- **Native Hawaiian or Other Pacific Islander** – "A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands."
- **White** – A person having origins in any of the original peoples of Europe, the Middle East, or North

defined by the US government at the start of this study (Bureau, 2022). This is the notion of race that underpinned the critical view of race in this study. This taxonomy lacks any meaning that might be important in public health research. Definitions are not tied to any meaningful shared

attributes. In the 2020 US census, 204,277,273 persons were classified as White (*United States - Census Bureau Profile*, n.d.). In addition to other drawbacks, this large and undefined group disadvantages White subpopulations with disproportionate health burdens (Bhopal & Donaldson, 1998). The US White population is equivalent to the population of Brazil (approximately 218,000,000) – the seventh largest country in the world (*Country Comparisons - Population*, n.d.).

Changes in racial classifications and data collection practices in the US census have shifted over time. A highly dramatic period in the US census occurred in the twenty-year span from 1890 to 1910: "Black" or "Mulatto" in 1880, "Black," "Mulatto," "Quadroon," or "Octoroon" in 1890, or "Black" in 1900 (Cohn, 2010; Nobles, 2000). In 1910, the use of "mulatto," was intended to distinguish between "full blooded negroes" and persons having some proportion of "Negro blood" (Cohn, 2010). The "one-drop rule," denoting any appearance of "Negro blood," was most characteristic of US censuses (Nobles, 2000).

Before 1960, census takers defined the race of census respondents rather than relying on self-identification. The 2000 US census marked the first time that more than one race could be selected, meaning that prior Census results underreported multiracial identities (T. N. Brown, 2003). There have been longstanding calls for "Hispanic or Latino" to be listed as a race (CDC, 1998). The United States Office of Management and Budget (OMB) published an announcement on adding Hispanic and Middle East/North African for federal data collection in March 2024 (*Revisions to OMB's Statistical Policy Directive No. 15*, 2024). This policy change is highly consistent with the historical instability of race.

For the first time in the history of the US census, respondents could provide an "origin" write-in under race for White and Black or African American in 2020. The US census describes

these open-ended responses as “origins”. It is worth noting, however, the bias of the US 2020 census form. First, ancestry or ethnicity is subsumed under race for White, Black or African American, or American Indian or Alaska Native — not for the two other groups. Although the US Office of Management and Budget (OMB) purports that there are five “races,” two of the five are not actually listed as such, meaning “Asian race” or “Native Hawaiian or Other Pacific Islander race” do not appear. Rather, options are presented alongside a write-in option: Chinese, Filipino, Asian Indian, Vietnamese, Korean, Japanese, Native Hawaiian, Samoan, and Chamorro.

Under White, Black or African American, or American Indian or Alaska Native, respondents can add an origin. Each group is provided with examples. For White, they are “German, Irish, English, Italian, Lebanese, Egyptian, etc.” For Black or African American, they are “African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.” American Indian or Alaska Native includes “Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.” Questions about origins have appeared previously in censuses, but separately from race (Wang, 2018).

“African American” is purported as a race alongside Black. It is also listed as an example of an origin. “African American” cannot be both a race and origin or ethnicity because race has been posited in the US census as a universal construct associated with Old World origins (e.g., Europe, Africa, Asia) – not that it is meaningful or accurate. Charles Gallagher, a sociology professor at La Salle University has noted another concern about origins for self-identified white persons, “If you have a population that's been in the U.S. for a very long time and people have been, you know, crossing the ethnic line and dating and marrying, people aren't going to have a real accurate record" (Wang, 2018) This discussion of the US census highlights many concerns with its implicit bias toward favoring the legitimacy of race and subsuming of ancestry, national

origin, and cultural identity under race for three groups. The issues should be met with much skepticism in science. 2024 marked the first time in twenty years for the minimum categories on race and ethnicity (*US Changes How It Categorizes People by Race and Ethnicity. It's the First Revision in 27 Years*, 2024).

III. Internal Validity and Race in Research

In the seminal text of Cook and Campbell, they defined internal validity as, "approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause" (Cook et al., 1979). In experimental studies, internal validity refers to the causal relationship between the treatment and outcome (Slack & Draugalis, 2001). A study with high internal validity has limited the exogenous influence on the effect of the independent variable(s) as the most likely cause of changes in the dependent variable (Flannelly et al., 2018). Alternate explanations do not account for the observable change in the dependent variable. Interval validity is inherently weaker in non- or quasi-experimental designs. Internal validity in non-experimental studies refers to the extent that the association of the exposure(s) to outcome(s) can be inferred (Matthay & Glymour, 2020). The inherency of the "race" threat to internal validity arises from poor reliability and validity, as well as other threats and biases. The following discussion of these threats is not intended to be exhaustive, rather to highlight critical areas.

Race is inherently marked by lack of reliability and often used as a proxy for any number of constructs such as biology, culture, or ancestry (Kittles et al., 2007; Schmidt et al., 2023). Some researchers have argued for race as a proxy for racism (Gee & Ford, 2011; D. R. Williams & Mohammed, 2013). However, race-as-proxy is imprecise and attenuated. Research studies that lack a robust theoretical framework may be prone to a several errors: Type-I (finding

significance where there is none), Type-II (not finding significance where there is), Type-III error (inaccurately determining the direction of difference for statistically significant groups).

Defined as omitting one or more important variables from the final model, mis-specification error often stems from a "weak or non-existent theoretical framework for building a statistical model" and leads to errors in research studies (Onwuegbuzie, 2000). "Mis-specification error is perhaps the most hidden threat to internal validity" (Onwuegbuzie, 2000). Measures of racism and multi-dimensional measures of cultural and racial identity may be more appropriate than a single-item survey tool to mitigate this error (Garcia et al., 2015; Johnson & Carter, 2020; Schmidt et al., 2023).

Internal validity is concerned with the effect of the independent variable(s) as the most likely cause in the change of the dependent variable. Race does not fall under the epidemiological definition of a cause on account of its assumed immutability, "(F)actors under consideration as potential causes must be plausibly manipulable; they cannot include fixed attributes such as race" (Kaufman & Cooper, 2001).

Race is manipulable in counterfactual research, "the existence of counterfactual distributions, such as the outcome distribution for Whites, had they been Black" (Kaufman & Cooper, 2001; Morgenstern, 1997). The counterfactual framework was applied in a study of mortality and racial disparities, "The direct effect is defined as how much mortality would kept at the level when race is White. The indirect effect is defined as how much mortality would change on average when race is Black, but the mediator was changed from the level it would take when race is White to the level when race is Black" (Luo et al., 2022). The Oaxaca-Blinder decomposition approach allows researchers to assess differences in outcomes if the same mean levels of variables are assumed between groups (Cuevas et al., 2020).

Racism is a key driver of racial health disparities and encouraged in health research as a substitute for race (Kaufman & Cooper, 2001; Mateo & Williams, 2021). Racism is widely accepted as common in the US and fits the epidemiological definition of an exposure, "any factor that may be associated with an outcome of interest." (CDC, 2021; Lee & Pickard, 2013; *NIH Stands against Structural Racism in Biomedical Research*, 2021). The Centers for Disease Control and Prevention described the impact of racism as, "pervasive and deeply embedded in our society—affecting where one lives, learns, works, worships and plays and creating inequities in access to a range of social and economic benefits—such as housing, education, wealth, and employment" (CDC, 2021). There are no fewer than 14 measures of perceived racism and discrimination (Atkins, 2014). Racism may be appropriate than race (Schmidt et al., 2023).

Race is associated with higher morbidity and mortality "across a wide range of health conditions, including diabetes, hypertension, obesity, asthma, and heart disease, when compared to their White counterparts" (CDC, 2021). Racism is pervasive and a major driver of social, health, and economic disparities (CDC, 2021). Racial inequity manifests in the wealth gap, residential segregation, employment, and education. As with the white-to-black family wealth gap, which is the same as it was in 1968, racial inequity is worsening in some cases (Khullar & Chokshi, 2018; McIntosh et al., 2020). Assuming that these trends matter for research, researchers cannot assume "responses of different cases are independent of each other," a core assumption in inferential statistics (Tabachnick et al., 2007). We later discuss arguments for clustering and nesting effects.

The National Institutes of Health (NIH) opined in March 2021, "racial injustices that have been allowed to endure over four centuries and that significantly disadvantage the lives of so many. The time for upholding our values and taking an active stance against racism, in all its

insidious forms, is long overdue" (*NIH Stands against Structural Racism in Biomedical Research*, 2021). Structural racism is defined as "the normalization and legitimization of an array of dynamics–historical, cultural, institutional and interpersonal–that routinely advantage White people while producing cumulative and chronic adverse outcomes for people of color" (Lawrence & Keleher, 2004). It is thought to be so common as be a "fundamental cause of persistent health disparities in the United States" (Churchwell et al., 2020). The study of structural racism is a recent, emerging field of study in the biomedical sciences.

Selection bias is also an area of concern in current research analysis and interpretation. Reliance on OMB's five federal racial groups (i.e., White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or other Pacific Islander) to establish that a sample is representative of a population of interest does not constitute strong scientific reasoning and poses inherent threats to internal validity and external validity. Control and experimental groups may differ from each other and a general population if assertions of representativeness only account for race.

Researcher bias is a threat to internal validity. "Researcher bias refers to researcher's tendency of having a partial perspective, which favors certain populations or opinions against the alternatives" (Gao, 2020). It can be introduced when i) research participants conform to the "perceived racial expectations of the (non-concordant) race of the interviewer," ii) researchers formulate the research question without input from the population of study, iii) introduce a biased conceptualization of race, or iv) make analytical decisions on the treatment of race (Bonham & Nathan, 2002; Davis, 1997). Racial categorization forms a type of researcher bias toward favoring the legitimacy of race in science, except that researchers often fail to explain

what construct race or ethnicity data collection is intending to capture (Kaufman & Cooper, 2001).

Racism is well-established as a determinant of health (Churchwell et al., 2020; *NIH Stands against Structural Racism in Biomedical Research*, 2021). This widely held assumption raises a key challenge to the assumption of independence required in most statistical tests. The assumption of independence means that observations are uncorrelated with each other (Liang & Zeger, 1993). "While some solutions to nonindependence exist at the statistical analysis stages, there is little advice on what to do when complex analyses are not possible, or when studies with non-independent experimental designs exist in the data" (Noble et al., 2017). Analytical adjustments may not be possible due to study design, sample size, and lack of statistical power.

Even if the assumption of independence can be met, considerable challenges arise to meet other statistical assumptions. If dichotomized racial categories (e.g., Black or African Americans compared to Whites) are analyzed using t-test or ANOVA, ANOVA results may be unreliable. The *F* test is robust to mild violations of normality, particularly when group sizes (e.g., racial groups) are equal (Troncoso Skidmore & Thompson, 2013). However, violation of the homogeneity of variance assumption tend to increase Type 1 error rate, an effect that is more pronounced in an unbalanced design (i.e., unequal group sizes)(Troncoso Skidmore & Thompson, 2013). If the smaller sized group has the larger variance in one-way between-subjects ANOVA, then the F test becomes "too liberal, leading to a Type I error rate (inflated alpha level)" (Tabachnick et al., 2007). Several inferential models (e.g., ANCOVA) assume that the slopes of the regression of the dependent variable and covariate (e.g., race) are equal.

Obvious clustering effects may be present in health disparities studies. This means that other methods for studying racial health disparities may be more appropriate such as

multidimensional scaling, multilevel or mixed effects models. “Thus, if, in the data analysis, a submitted article ignores obvious clustering that needs to be captured or considered, the statistical editors will ask for justification of this or for a reanalysis accounting for clustering using an approach suitable for the estimand of interest…A multilevel or mixed effects model might be recommended, for example, as this allows cluster specific baseline risks to be accounted for and enables between cluster heterogeneity in the effect of interest to be examined.” (Riley et al., 2022). Ignoring clusters can also cause overestimation of random effects (Park & Chung, 2022). Races are generally not considered statistical cluster groups. If race is a proxy for levels of exposure to racism based on structural racism theory, then clustering is likely a more accurate theoretical and scientific basis for analysis than conducting statistical tests that assume independent, unrelated subjects. Still, racial “membership” and racism and are not synonymous conceptually or operationally.

Alternatively, nesting is defined as the organization of data at more than one level. Nesting effects occur from natural groups and have been identified in research studies at the level of the family, household, school, neighborhood, state, and hospital (Peugh, 2010; Tabachnick et al., 2007). Racial health disparities research may be explained by a racial level of predictor. The case for nesting in racial health disparities research seems plausible. If not accounted for when present, “the Type I error is inflated because analyses are based on too many degrees of freedom that are not truly independent” (Tabachnick et al., 2007). A common example that appears in the literature involves a school-based intervention in which classrooms are randomly assigned different intervention methods such that students are nested in classrooms (Tabachnick et al., 2007). In analysis, individual student outcomes would be a lower-level predictor while classrooms would be higher-level. Nesting effects also appear in a cross-

sectional population study when multiple responses come from the same household or treatment providers in clinical studies (Wampold & Serlin, 2000).

Race is commonly used as a confounding or "control" variable in research because it is not considered a primary variable, "merely a nuisance in the data" (Kaufman & Cooper, 2001; T. A. LaVeist, 2000). Controlling confounding effects can diminish power, reduce external validity, and increase variance, and improve accuracy and the ability to assess intervention effects (Ferguson, 2004; Greenland et al., 2016). Adjusting for race can carry risks by overcontrolling and biasing estimates of total effects. Unmeasured confounders also pose issues in research – producing over- or underestimates of the direct and indirect effects. Mis-specification error can contribute to unmeasured confounding. In frameworks that use race as a proxy for socioeconomic status, adjusting for the confounding effects of socioeconomic status would bias the statistic for race (Shavers & Brown, 2002).

Researchers frequently collapse racial categories during analysis due to small sample sizes or unbalanced design (i.e., unequal group sizes). This can result in increased Type 1 error rates, incorrect inferences, lack of model fit, and under- or over-inflated power (Rutkowski et al., 2019; Strömberg, 1996). Decisions about not retaining the original collection data should be theoretically justified while avoiding discouraged practices such as "non-white" categories (Flanagin, Frey, Christiansen, & Bauchner, 2021).

IV. External Validity and Race in Research

External validity refers to the degree to which the study findings are generalizable to what populations under what conditions (Slack & Draugalis, 2001). In the classic text of Cook and Campbell, they defined external validity as, "approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the

absence of cause which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times" (Cook et al., 1979). No fewer than twelve threats to external validity are present at the design and data collection stage and five at the data analysis stage (Onwuegbuzie & Daniel, 2003). Our discussion highlights three major threats to external validity: population validity, ecological validity, and temporal validity that are important to the racial threat to research quality.

Population validity. Population validity refers to how well study findings can be generalized to larger target populations and subgroups within those populations (Onwuegbuzie, 2000). Study participants' race is often used to assess representativeness of the analytic sample to the total sample and target population. Racial representativeness assumes monolithic, fixed racial groups in most biomedical research studies despite genetic, historical, linguistic, cultural, geographic, and religious diversity within and among racial groups. Study conclusions on the difference or risk for one group over another has population validity to the extent that populations are conceptually well-defined. The hypothetical phrase, "Black people have six times greater risk of *x-disease* from White people" is interpretable to the extent that the study clearly and accurately defines "Black people" and "White people". It is rarely the case that studies have high interpretability in such an example. Analytical decisions may reduce generalizability. Researchers are discouraged from using terms such as "non-White" or "other", (Flanagin, Frey, Christiansen, & AMA Manual of Style Committee, 2021). Collapsing of data reduces its scientific value and practical significance.

Ecological validity. Ecological validity concerns the generalizability of study findings across settings, contexts, and conditions (Christensen & Waraczynski, 1988; Onwuegbuzie & Daniel, 2003). Researchers would need to address how race holds up across settings, contexts,

and conditions – a tall order. We further emphasize this point by drawing from the gentrification health literature.

The neighborhood influence on health is a determinant of health that can undermine ecological validity. Black and low-income populations tend to have negative health impacts associated with neighborhood change (Smith et al., 2017). However, gentrification, defined as an acute form of neighborhood change associated with Black displacement and cultural displacement, tends to be concentrated in certain neighborhoods within half a dozen US cities (Gibbons & Barton, 2016; Tulier ME et al., 2019; C. Williams, Woodard, et al., 2022). The degree of neighborhood spatial social and economic polarization between high-income White and low-income Blacks has shown mixed findings in the literature: lower odds (e.g., hypertension) versus higher odds (preterm births, infant mortality, pregnancy-associated mortality) (Chambers et al., 2019; Dyer et al., 2022; Feldman et al., 2015).

In other words, varying intensity in neighborhood change can cause differences in effects within certain populations. We underscored a key problem with relying on race to establish ecological validity because we cannot assume that racial experiences are inherently generalizable – ecologically or temporally.

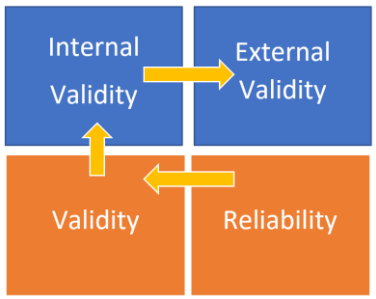
Temporal validity. Temporal validity refers to whether study findings can hold across time. As discussed, the meaning of race is apt to political and social re-construction. Social movements, governmental policies, and judicial decisions can catalyze epochal changes to the meaning of race. The decision of the US government to allow more than one race to be selected in the 2000 US Census is an example of a shift in temporal validity. The decision of the US federal government to add new monolithic races or ethnicities for Middle Eastern/North African and Hispanic is another example (*Revisions to OMB's Statistical Policy Directive No. 15*, 2024).

We later discuss the role of the United States Office of Management and Budget (OMB) in setting federal standards on race data collection and analysis that impact public health datasets.

3. Conceptual Frameworks

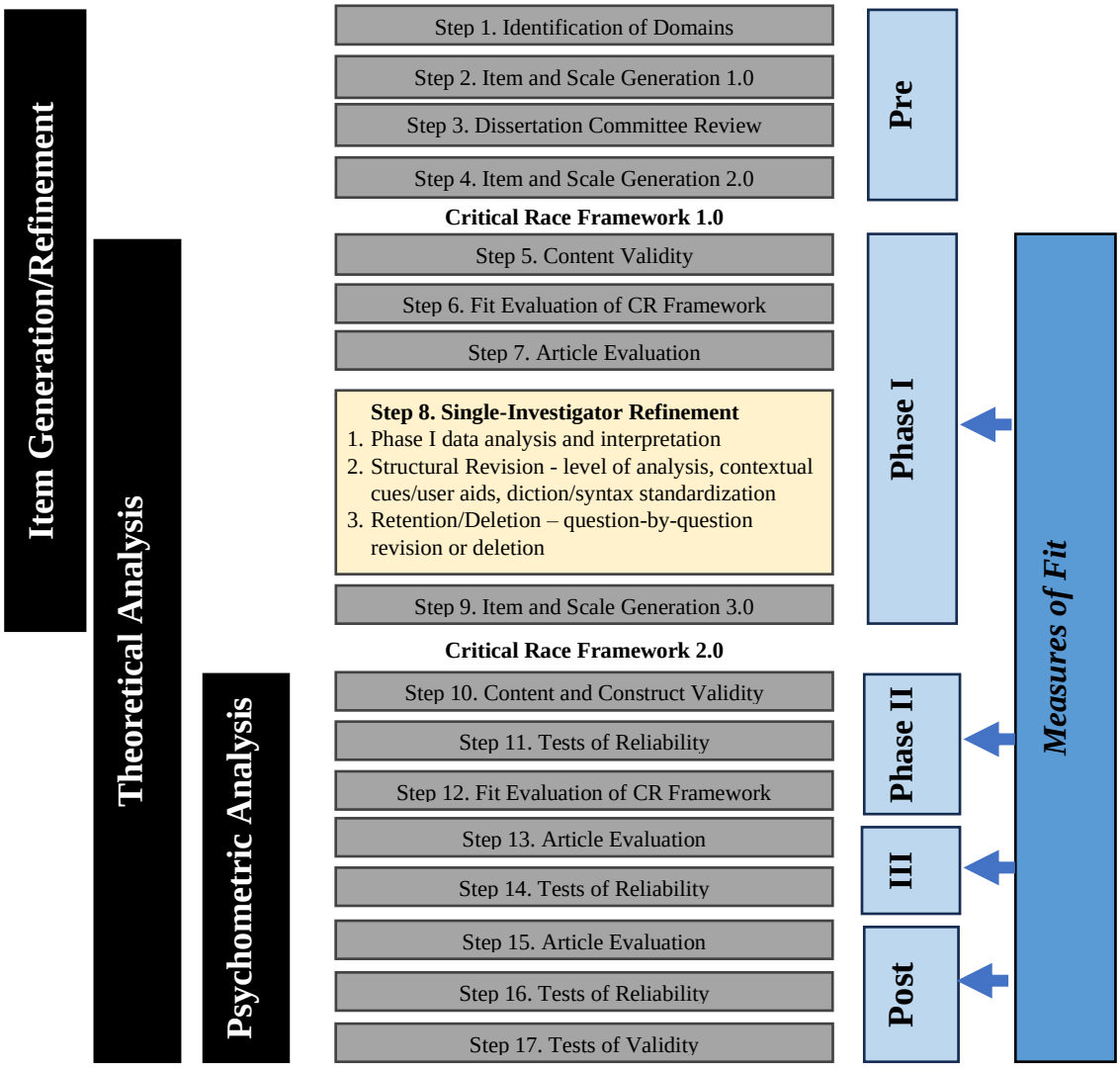
Theoretical Framework for Critical Appraisal. The conceptual model focuses on the quality of the race by first evaluating the data collection tool (reliability and validity), then the strength of causality or association (internal validity) and generalizability (external validity) (Figure 1). The importance of these concepts to judge the quality of a study is well-established. We acknowledge that our heuristic model may be considered an oversimplification of concepts within critical appraisal theory. However, the Critical Race Framework is not dissimilar to the structure and content of critical appraisal tools that appear in the literature, as we discuss further in Chapter 5. This conceptual framework was essential for organizing questions in the Critical Race Framework.

Figure 1
Conceptual Framework for Critical Appraisal



The Conceptual Framework for the Critical Race Framework Development. We adopted a conceptual model to guide development of the Critical Race Framework. *The Conceptual Framework for the Critical Race Framework Development* is a 17-step expert-engaged and PI-initiated framework (Figure 2). We borrowed from the model of Morgado and colleagues to define three basic steps: 1) instrument and scale generation to include evaluation of scale appropriateness, instructions, and question quality, 2) theoretical analysis to assess content validity, and 3) psychometric analysis to establish construct validity and reliability (Morgado et al., 2017). We adopted components of the scale development model by Boateng and colleagues (Boateng et al., 2018). The 17-step framework is tailored to this study to reflect its delimitations,

Figure 2 – Conceptual Framework for Critical Race Framework Development



such as the lack of an expert panel in initial item generation. In addition, we added *measures of fit* — acceptability, appropriateness, feasibility, satisfaction, relevance, and audience and setting appropriateness. These components are discussed more fully in Chapter 3.

Reliability. Many factors influence reliability such as source of information, categorization, survey features, study conditions, and conceptual meaning (Flanagin, Frey, Christiansen, & Bauchner, 2021). Inadequate testing of these sources can cause systematic bias, meaning that findings cannot be reliably or accurately interpreted without bias corrections. In the

context of this study, we focus on critical areas of reliability theory. Table 3 lists our scope of inquiry for reliability theory (Table 3). Researchers rely on experts' quantitative and qualitative judgment to determine the quality of reliability evidence. It is common for critical appraisal tools to gather data on reliability over several years. For example, the Mixed Methods Appraisal Tool was first developed in 2006, then revised in 2011 and 2018 (Hong et al., 2019; Pace et al., 2012; Pluye et al., 2009).

Critical appraisal tool development relies on *expertise* – specialists with high familiarity with relevant theoretical concepts and assessments. These experts can be recruited from within a professional network or other sampling methods. They use expert judgment throughout appraisal stages. Structured approaches to reach consensus or define content (e.g. Delphi techniques) during initial stages are common. The role of affective states is widely researched in a range of disciplines and settings (Ainley, 2006). We defined *affect* in this study as the emotions involved with adaptive appraisal and behaviors (e.g., enrollment and attrition). Experts are not expected to completely agree in any stage of critical appraisal tool development and testing. It is necessary to explore whether study findings are attributable to affective states.

Study bias is a major consideration for any new tool. Bias can arise from the tool or survey itself and affect the accuracy and reproducibility of data. Researchers should be highly sensitive to this area of reliability. Many *behaviors* are observable (e.g., straight-lining, drop-out) and may attenuate reliability. Data analysis and interpretation (e.g., choosing tests, missingness, imputation, collapsing) constitute our final two areas of reliability analysis. At each stage, decisions should be explained and justified. Decisions to improve internal validity can limit external validity.

Interrater reliability (IRR) demonstrates the robustness of a scale to different raters (Bobak et al., 2018). Technically, IRR refers to cases of ordinal or interval data (O'Connor & Joffe, 2020). Interrater agreement (IRA) indices assess the extent to which two or more raters are concordant. IRA indices focus on identical scores among rater in contrast to IRR which accounts for measurement error (Gisev et al., 2013). IRR and IRA are necessary components of reliability and validity. Our conceptual model includes them under reliability.

Table 3: Realms of Inquiry in Reliability Theory for the Critical Race Framework Study: Areas of Focus

1. **Expertise** – Specialized knowledge of and familiarity with theoretical concepts.
2. **Judgment** – Applied knowledge and qualitative judgment.
3. **Affective** – Feelings or emotions influencing behaviors (e.g., attrition) and responses.
4. **Experiential/Attributes** – Research and training factors that influence survey data.
5. **Study bias –** The biasing effects of the study design or instrument on estimates.
6. **Behaviors** – Behaviors that may lead to inaccuracy (e.g., drop-out, straight-lining).
7. **Data Analysis –** Decisions (e.g., imputation) and appropriate tests (e.g., correlation).
    - *Interrater Reliability* – Agreement or consistency among raters.
8. **Data Interpretation –** Accurate data interpretation considering reliability evidence.

Validity. Table 4 defines the critical areas of validity theory in this study. As with reliability, validity relies on expert qualitive judgment and appropriation of relevant test. Content validity is defined by a qualitative and empirical process to judge whether items reflect a given domain in new scale and instrument development. Determining whether measures are theoretically linked is usually based on expert qualitative evaluation (Weiner et al., 2017). Four major domains of content validity appear in the literature – "domain definition, domain representation, domain relevance, and appropriateness of test construction procedures" (Almanasreh et al., 2019). Doman definition refers to the quality of operationalization of a particular construct (Almanasreh et al., 2019). Domain representation concerns how well a tool represents and measures a construct of interest (Almanasreh et al., 2019). Domain relevance

addresses the relevance of item within an instrument to a related domain (Almanasreh et al., 2019). Appropriateness of procedures involves assessment of the procedures in the construction of the instrument to ensure representatives and relevance (Almanasreh et al., 2019).

Table 4: Realms of Inquiry in Validity Theory for the Critical Race Framework Study

1. **Content Validity** – Expert-informed development and validation of items that reflect a given domain. In this study, we mean the four core constructs of the CR Framework.
2. **Construct Validity –** The evidence-based reasoning among experts to establish that a measurement tool accurately represents a construct. In this study, we are concerned with *measures of fit*: acceptability, feasibility, and appropriateness.

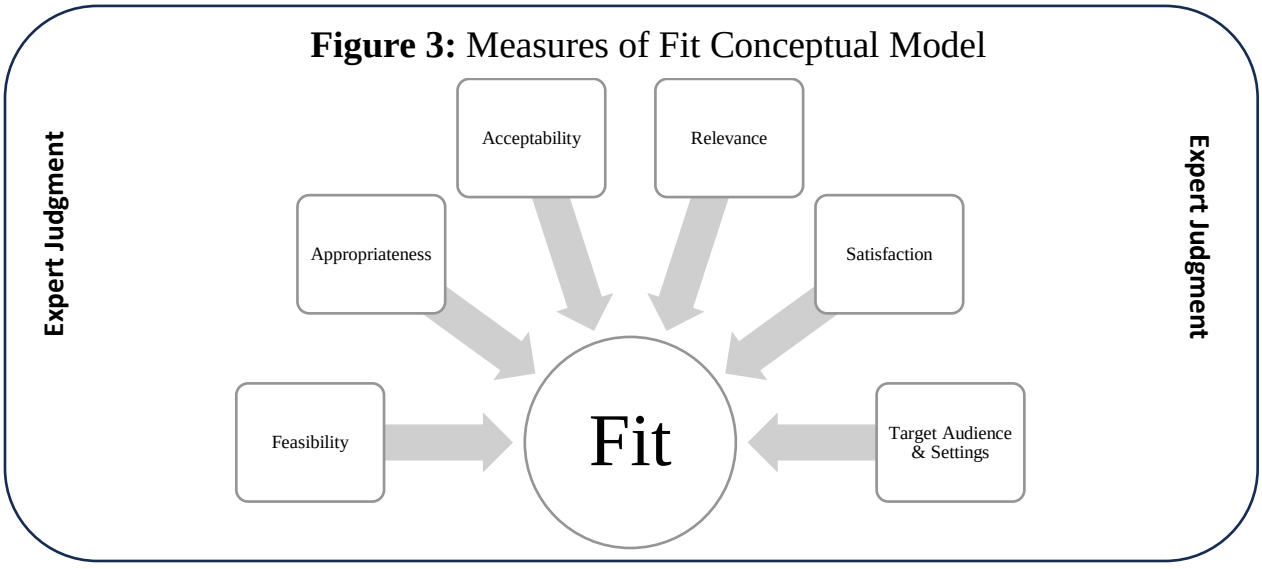
Appropriate selection of an index depends on the number of raters, study variables of interest, and the context of the study (Gisev et al., 2013). ICC selection "involves identification of the type of reliability study to be conducted, followed by determining the "Model," "Type," and "Definition" selection to be used" (Koo & Li, 2016). There are several versions of the ICC for interrater reliability that warrant careful consideration based on the model, consistency or absolute agreement, and k- or single-raters measurement (Bobak et al., 2018; Koo & Li, 2016).

Measures of Fit. Acceptability, appropriateness, and feasibility are constructs in research innovation used to evaluate fit or match of a new tool or practice (Weiner et al., 2017). These *measures of fit* are important indications of potential implementation effectiveness. We adopted the framework of Proctor and colleagues, which was intended for treatment implementation and clinical outcomes (Proctor et al., 2011; Weiner et al., 2017). We avoided five of the eight aspects of their model due to concerns about survey fatigue, construct redundancy, and different settings for implementation. Theirs was a clinical setting. Within the context of this study, acceptability is the perception among raters with high educational attainment in a public health specialty that the Critical Race Framework is agreeable, palatable, or satisfactory. Appropriateness refers to

relevance, fit, and compatibility of the Critical Race Framework. Feasibility concerns the extent to which the Critical Race Framework can be successfully used.

Our conceptual model supplemented *fit* with satisfaction, relevance of the instrument to key core concepts of critical appraisal, and appropriateness to public health audiences and settings (e.g., researchers, students, classroom) (Figure 3). Satisfaction was needed because it captures a larger concept of general nature – that of experience with the Framework. We ensured that in operationalizing satisfaction that we avoided survey burden. We also needed to establish for whom, beyond our eligible population, could the CR Framework be appropriate and targeted. We assumed that a tool of highly relevant items would enhance user experience and uptake. This model limits evaluative perceptions to public health experts. The lowest level of educational attainment is doctoral student or master's degree (Phase I) or terminal degree in public health (Phase II).

Figure 3: Measures of Fit Conceptual Model



Theory-Informed Web-Based Training. Andragogy is the study of adult learning and education. It contrasts with pedagogy, which is concerned with children's education. Malcolm Knowles first introduced the distinction between adult and child learning in the United States (US) in the 1970s (Knowles et al., 2014). The classic framework for andragogy concerns six assumptions about

adult learners: 1) They need to know the practical import of instruction and content (why they are learning); 2) A learner self-conceptualizes as autonomous and self-directing; 3) A learner relies on prior experiences and knowledge, 4) a learner exhibits readiness to learn; 5) A learner is oriented to learning that is problem-centered and content-dependent; 6) A learner is motivated to learn drawing from intrinsic drive and anticipation of payoff (Knowles, 1980, 1984; Knowles et al., 2014). Adult learners also seek to be involved in the content development and evaluation process (Bryan et al., 2009). These assumptions are shaped by individual and situational differences, purposes for learning, and institutional and societal growth (Knowles et al., 2014).

Adult learning theory has been translated for public health practice as: defining job responsibilities within organizational goals, specifying the goals of training to improve an existing program component, conducting pre-assessment to gain insight into learner motivation and orientation to lifelong learning, conducting inventory of learners' prior experience on which to capitalize for training, and involving learners as stakeholders as curricular builders (Bryan et al., 2009). We relied on adult learning theory in developing the web-based training.

4. Use of Race in Behavioral Health

This dissertation study has major implications across public health specialties and practices. However, this study arises within a department of behavioral and community health – the University of Maryland School of Public Health Department of Behavioral and Community Health. Thus, a disciplinary focus in behavioral health is warranted. Race is a common variable in health behavioral science. Racial categorizations consistently involve evaluating differences in health behaviors and risk factors between non-Hispanic black and white populations (Andersen et al., 1998; Dubay & Lebrun, 2012; Mezuk et al., 2010; Otten et al., 1990; Redmond et al.,

2011; Thorpe Jr et al., 2015). A disciplinary review was conducted to support the premise of this study for behavioral health science.

We conducted a Google Scholar search on January 12, 2023, for "health behaviors" and "racial disparities" between 2018-2022 and retained the first ten results. In total, these articles had 174 citations. None of the ten studies defined what construct race was intending to capture (Bell et al., 2018; Cuevas et al., 2020; Devaraj et al., 2020; Haq & Penning, 2020; Luo et al., 2022; McClendon et al., 2021; Murray Horwitz et al., 2018; Parcha et al., 2020; Tuthill et al., 2020; Whitaker et al., 2018). All studies assumed for the most part that a “true value” of race existed. No study questioned the scientific validity of race variables. The majority of studies primarily examined non-Hispanic Blacks and Whites using self-reporting (Bell et al., 2018; Devaraj et al., 2020; Luo et al., 2022; McClendon et al., 2021; Parcha et al., 2020; Whitaker et al., 2018). Hispanic status was collected and analyzed for the remaining studies except for one (Cuevas et al., 2020; Murray Horwitz et al., 2018; Tuthill et al., 2020). A Canadian study examining social determinants and cognitive functioning was the only study from our results to collect a wide variety of racial and ethnic categories (e.g., White, Chinese, South Asian, Black, Filipino, Latin American, Southeast Asian, Arab, West Asian, Japanese, Korean)(Haq & Penning, 2020). It was also the only study to collapse the diverse analytic sample into White v. non-White, a practice that is strongly discouraged (Flanagin, Frey, Christiansen, & Bauchner, 2021). As discussed, this can result in increased Type 1 error rates, incorrect inferences, lack of model fit, and under- or over-inflated power (Rutkowski et al., 2019; Strömberg, 1996).

None of the studies discussed potential threats to internal validity due to a lack of racial conceptualization. During the data collection phase, study methods varied considerably. None acknowledged potential errors with collecting race data except a single study that sought to

"verify" race. In their study of cardiovascular health behaviors, Whitaker and colleagues "verified (race) at the first and second exam," then used the second exam to determine the "confirmed race" for analysis (Whitaker et al., 2018). The authors do not elaborate on this confirmation process (Whitaker et al., 2018).

Tuthill's study on race and sexual identity in health behaviors initially allowed participants to self-report multiple races, then forced self-reporting of a single race in a follow-up ("follow up question asking which group best represents the respondent's race") (Tuthill et al., 2020). As we discuss later, the Behavioral Risk Factor Surveillance System (BRFSS) uses a similar series of questions on race. Cuevas and colleagues relied on the Chicago Community Adult Health Study (2001-2003) – an in-person cross-sectional study involving 3,105 adults in 343 neighborhood clusters and multiple survey teams (Cuevas et al., 2020).

All ten studies fully assumed that race was an appropriate scientific variable. None questioned the inherent threats to internal and external validity that race posed. In fact, several studies could have introduced biases and increased type I or II errors due to decisions during analysis — collapsing all non-white respondents, excluding respondents with multiple race, excluding respondents who selected "other," and combining "other" race with White (Cuevas et al., 2020; Haq & Penning, 2020, 2020; McClendon et al., 2021, 2021; Murray Horwitz et al., 2018). These threats to internal and external validity are underexplored in all studies.

These findings support the premise of this study on the under-reporting of theory and methods of racial variables. This brief research report highlighted weaknesses in conceptualization, data collection, and analytical treatment of race for ten behavioral health studies. These findings may not be generalizable to the field of behavioral health, but they underscore the significance of our study.

5. Review of the Literature – Critical Appraisal Tools

We conducted literature reviews to determine a) whether a bias tool akin to the CR Framework existed and b) to identify critical appraisal tools to which we could compare ours. To answer the former, we conducted non-systematic and systematic searches. Prior to Phase I, we performed a non-systematic search in PubMed, Web of Science, and the University of Maryland OCLC database using variations of the following terms: racial measure, framework, article critique, structured review, and article analysis on December 5, 2022. The results yielded no results. A subsequent systematic search occurred prior to Phase II in July 2023 to identify any similar bias tool that may have been published after December 2022. A list of synonyms was first compiled based on standard English thesauruses and a review of related terms used in the literature: "appraise," "evaluate," "critique," and "analyze." Our methods are described below.

Systematic Search for Critical Appraisal Tools for Use of Race in Research. The first search utilized MEDLINE, Scopus, and Web of Science Core Collection were systematically searched from inception involved a titles-first approach. In this approach, the title of each article is screened, then the abstracts of the remaining results are reviewed. As opposed to a concurrent review of the title and abstract, the title-first approach has shown time and cost efficiency (Mateen et al., 2013). Article were retained if they specifically provided a bias or critical assessment tool with respect to critical perspectives on the use of race in research.

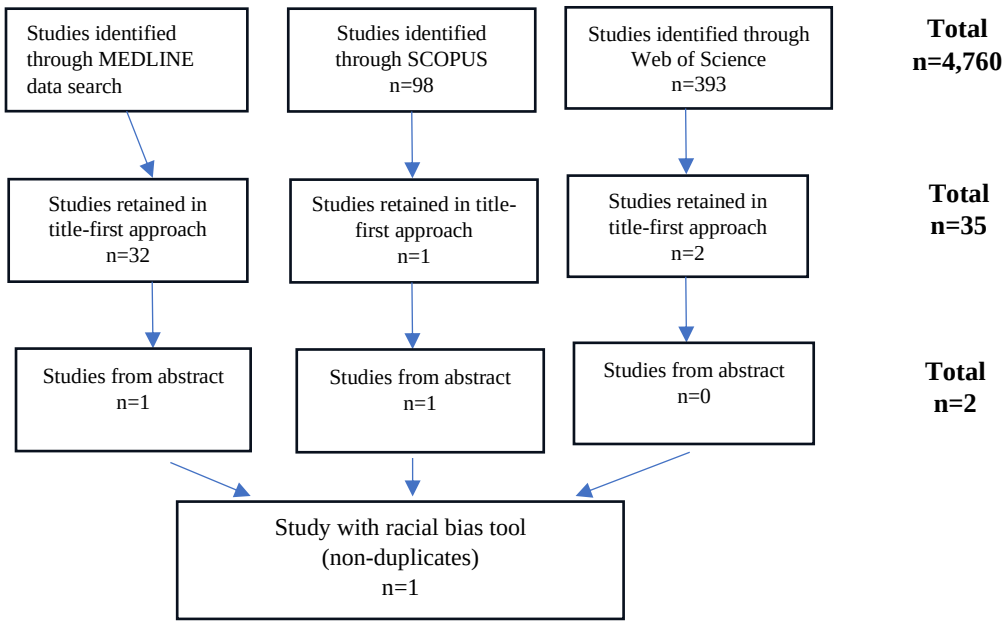
The search in MEDLINE and Scopus was operationalized using Boolean expression as follows: appraise the use of race OR evaluate the use of race OR critique the use of race OR analyze the use of race. The search was open to all articles published in the MEDLINE database from inception (1975 – present). The search in Web of Science was limited to highly cited papers

and abstract since the initial search generated 52,694, which would have exceeded available resources to review.

A limited, non-systematic search of critical appraisal tools was also conducted to inform refinement of the Critical Race Framework 1.0, meaning the transition between Phase I and Phase II. Results from the 10 most cited critical appraisal tools and related terms within the last six years (2018-2023) were used for comparison in the discussion of study results.

The total search yielded 4,760 results. Medline had 4269 total results of which 32 were retained from a title-first approach and a single article was retained from an abstract review (Figure 4). Scopus had 98 results of which a single article (1) was retained from a title-first approach and a single article (1) was retained from an abstract review. Web of Science had 393 results. Two were retained from a title-first approach and none from an abstract review.

Figure 4: Results from Database Searches: Medline, Scopus, and Web of Science to Identify Bias Tools Similar to the Critical Race Framework



The only tool similar to the Critical Race Framework to appear in the literature is the Critical Appraisal of Race in Medical Literature (CARMeL) tool (Garvey et al., 2022). The

CARMeL tool is intended for undergraduate (UME) and graduate medical education (GME) learners to critically reflect upon the use of race in biomedical literature (Garvey et al., 2022). It includes a 90-minute workshop with facilitator and participant guides. It provides a flow chart to guide learners by first identifying whether the construct of race is biologic or sociopolitical. Learners can reach one of four conclusions: 1) "Race is used as a sociopolitical construct with appropriate methods and no significant threats to internal or external validity," 2) "Race is used as a sociopolitical construct, but with some threats to internal or external validity," warranting methodological changes, 3) "Race is likely used as a biologic construct, but with few to little other threats to the internal validity of the study, "yield(ing) accurate observations of racial inequities in health, but draw inaccurate conclusions regarding causality, and 4) "Race is used as a biologic construct with significant threats to internal and/or external validity," justifying the learner to "discard" the study (Garvey et al., 2022).

The primary feature of the CARMeL Tool delineates between sociopolitical and biologic interpretation of race in research. The sociopolitical delineation pertains to structural inequities that may explain scientific findings. Whereas biologic explanations assume innate racial differences. This tool largely supports current practices in the use of the race in research insofar as research avoids biologic generalizations. It offers no evidence whether sociopolitical differences exist or if such differences are meaningful. Mere assertion of sociopolitical divides equates to a faulty premise.

The authors' reasoning is insufficient because it 1) lacks a clear sociopolitical conceptual framework, 2) does not rely on convergent and divergent validity, 3) does not address weaknesses in reliability, and 4) does not address limited external validity. The CARMeL tool also does not define their preferred set of racial groups and whether ethnicity, nationality, or

culture matter, if at all. The authors of the CARMeL did not fully situate support of sociopolitical race in scientific reasoning and evaluation.

Search for Critical Appraisal Tools. A search for "critical appraisal tool" AND “health” in Google Scholar and PubMed were conducted on July 26, 2023. Since this is not intended to be a comprehensive search, the five most cited studies were retained, respectively. The seven tools are compared to the Critical Race Framework during the study discussion. The search criteria were limited to the last six years (2018-2023).

The Google Scholar search resulted in 18,4000 results and was limited to the five pages of the Google search results. The results are organized by best match. Each page had up to 10 results. From among 32 total results, the five most cited studied had considerable range in individual citations – between 417 to 1510 – as of our search in July 2023 (Hong et al., 2019; Hong, Fàbregues, et al., 2018, 2018; Hong, Gonzalez-Reyes, et al., 2018; Munn et al., 2020). They included four studies by the same leading author pertaining to the mixed methods appraisal tool (MMAT) (Hong et al., 2019; Hong, Fàbregues, et al., 2018, 2018; Hong, Gonzalez-Reyes, et al., 2018). The fifth result concerned a critical analysis tool for evaluation the methodological quality of case series studies – the Joanna Briggs Institute (JBI) critical appraisal tool for case series (Munn et al., 2020).

The PubMed search produced 404 results. Our selection only drew from the first five search results. Each article was reviewed to identify the critical appraisal tool used in the study. This resulted in the following tools: Critical Appraisal Tool for Health Promotion and Prevention Reviews (CAT HPPR), JBI critical appraisal tool, Standard Quality Assessment Criteria (QualSyst) critical appraisal tool, Effective Public Health Practice Project critical appraisal tool,

Table 5: Summary of Critical Appraisal Tools Retails from Literature Search					
Name/Original Manuscript	**Purpose**	**Scale (latest version)**	**Reliability Testing**	**Validity Testing**	**Addressed Threats to Research Quality Due to Race Conceptualization, Collection, and Analysis?**
Mixed methods appraisal tool (MMAT)(Pluye et al., 2009)	Evaluate mixed studies – qualitative, quantitative, and mixed methods	Yes, No, Can' t tell	Interrater reliability was moderate to perfect. Item Kappa statistic varied from -0.17 to 1 ((Pace et al., 2012)	Agreement defined as consensus (score ≥ 0.80) was achieved for 7 out of 21 items.	No.
JBI critical appraisal tool for case series1	Evaluation methodological quality of case series - non-experimental study designs	Yes, No, Unclear, not applicable	—	—	No.
Critical Appraisal Tool for Health Promotion and Prevention Reviews (CAT HPPR)(Heise et al., 2022)	Evaluate scientific quality for systematic reviews, rapid reviews, and scoping reviews	Yes, No, NA	—	—	No.
JBI critical appraisal tool for quasi-experimental studies	Assess methodological quality of quasi-experimental studies	Yes, No, Unclear, Not applicable	—	—	No.
Standard Quality Assessment Criteria (Kmet et al., 2004)	Evaluate a variety of study designs using criterion-based approach	Yes, No, Partial, NA	Interrater agreement in quantitative and qualitative studies	—	No.
Effective Public Health Practice Project critical appraisal tool (Thomas et al., 2004)	Conduct systematic review for public health practice	Weak, Moderate, Strong	Intrarater reliability (Kappa statistic was 0.74 and 0.61)(Thomas et al., 2004)	Content validity Construct validity	No.
A critical appraisal tool for studies about Test and Trace programs	To screen COVID-19 modeling studies for quality and relevance	None	—	—	No.

Note: Dashes indicate that we did not find direct testing of reliability and validity. We do not purported that absent testing or reporting that there is no reliability or validity. Reliability and validity rely on quantitative and qualitative judgment. For example, relying on published tools to innovate a critical appraisal tool or approach supports validity, as does the use of experts in development. Every study appeared to rely on face validity, even if they did not test or report content validity.

and a critical appraisal tool for studies about Test and Trace programs (Fernández-Férez et al., 2021; J. Frank et al., 2021; Heise et al., 2022; Speyer et al., 2018; Xavier et al., 2018). JBI has produced no fewer than 15 critical appraisal tools (*JBI Critical Appraisal Tools | JBI*, n.d.). We retained seven tools between our searches. Table 5 summarizes each tool, including reliability and validity testing. None of the tools discuss race directly in critical appraisal.

6. Seminal Works in Research Quality: Is Race Covered?

As part of our literature review, we sought to determine whether our research premise had been covered in seminal works in internal validity and external validity. We conducted a Perplexity search on January 17, 2024 to identify a list of seminal texts using the search prompt, "What are seminal texts in internal validity and external validity in research" (*Perplexity*, n.d.). Seven works resulted: Campbell and Stanley (1963), Cook and Campbell (1979), Cook, Campbell, and Shadish, (2002), Trochim (2007), Onwuegbuzie (2000), Allen and colleagues (2011) (Allen et al., 2011; D. T. Campbell & Stanley, 2015; Cook et al., 1979, 2002; Onwuegbuzie & Daniel, 2003; W. Trochim & Donnelly, 2006; W. M. K. Trochim, n.d.-a). These results were largely consistent with our views of seminal texts on this subject.

We conducted keyword searches ("race" or "racial") and found that racial analysis is not considered a threat to research quality in these seminal texts. When Campbell and Stanley considered the "attitudes of whites who happened to be assigned to racially mixed versus all white combat infantry units," they implored researchers to "seek supplementary data to rule out plausible rival hypotheses, keeping aware of the remaining sources of invalidity" (D. T. Campbell & Stanley, 2015).

In addition to implying homogeneity and inherency, they did not consider the construct of race as a rival hypothesis. Similarly, Cook and Campbell relied on racial demographics to

support external validity, “If the percentages differ from what is expected on the basis of the (hopefully, recent) census, it would not be false to say that there is probably a race bias” (Cook et al., 1979). Trochim did not question the generalizability of race, equating it with age and sex, “In which categories of persons can a cause-effect relationship be generalized? Can it be generalized beyond the groups used to establish the initial relationship— to various racial, social, geographical, age, sex, or personality groups?” (W. Trochim & Donnelly, 2006). On quota sampling, “The problem here (as in much purposive sampling) is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education, race, or religion, etc.?” (W. M. K. Trochim, n.d.-b). These findings are not surprising given that contemporary norms and practices are an outgrowth of classical theories and approaches. The significance of the Critical Race Framework is underscored by these findings.

7. Behavioral Risk Factor Surveillance System: Source of Bias

We critically examined the Behavioral Risk Factor Surveillance System (BFSS) to underscore the severely underestimated threats to research quality consistent with our study premise. BRFSS is the premier behavioral health data program in the US. Established in 1984, it collects data on self-reported behavioral risk data, chronic disease, and utilization of preventive services in all US states, the District of Columbia, and three U.S. territories. More than 400,000 adults participate each year (*CDC - About BRFSS*, 2019).

BRFSS asks survey respondents to provide their race(s), “Which one or more of the following would you say is your race?” (Table 6). If respondents select more than one race, then they are prompted to select a single race, “Which one of these groups would you say best represents your race?”. The racial categories vary in acknowledgement of racial heterogeneity. “White,” “Black or African American,” and “American Indian or Alaska Native” do not have

subgroups. Whereas, “Asian” has seven coding subgroups. “Pacific Islander” has four. These are two-step patterns that we see consistently in many federal surveys, including the US census. The CDC does not make raw data on racial selection publicly available on its website.

The *Reactions to Race* module also asks about racial self-identification in a series of question about race-based treatment (Table 6) (Srivastav et al., 2021). States elect to use this optional module. “In 2022, the Reactions to Race Module was used with data collected from more states (28 states and the District of Columbia)” (*Statistical Brief on the Reactions to Race Module, Behavioral Risk Factor Surveillance System, 2022*, n.d.) After BRFSS survey administrators state the prologue protocol (“Earlier I asked you to self-identify your race. Now I will ask you how other people identify you and treat you”), they ask, “How do other people usually classify you in this country? Would you say: White, Black or African American, Hispanic or Latino, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, or some other group?” No subgroups appear in these response options (e.g., Chinese under Asian) as with the previous question on race. No explanation is provided for this difference in data collection compared to the prior question on rate. BRFSS provides no data on racial definitions or conceptualizations on its website or in the user guide. Race is imputed for respondents who refused to answer. The listing of “White” first, as opposed to alphabetically, might suggest racial normativity. The imputed values for missing race response derive from the most common race for the respondent’s state region. BRFSS generates variables for combined race and ethnicity (e.g., Black, Non-Hispanic) and race alone.

The inherent threat of racial nomenclature to research quality has not garnered critical appraisal even in reliability and validity assessments of BRFSS. Pierannunzi and colleagues’ systematic review of publications on reliability and validity of BRFSS identified thirty-two

studies from 2004-2011 (Pierannunzi et al., 2013). Their discussion did not identify weaknesses in data quality due to race conceptualization and analysis. This omission is not altogether surprising given the surfeit of evidence on this common occurrence.

BRFSS suffers from flaws in research methodology and analysis. First, implied homogeneity in "White," "Black or African American," and "American Indian or Alaska Native" populations is not supported by evidence. Arguments in favor of racial homogeneity within these populations are highly attenuated. The BRFSS underlying assumption on racial homogeneity constitutes a major source of research bias. Second, the social phenomenon of racial switching is common (Agadjanian, 2022; Agadjanian & Lacy, 2021). However, it is not considered a source of bias. Random or systematic bias is highly underestimated in BRFSS analytical approaches. Third, the BRFSS question on "best represents" introduces survey bias because this is only forced upon individuals with multiracial identities. A confirmatory process for racial identity of all respondents would be more methodologically justified, while avoiding singular racial identification. "Using 'best represents' race reduces the number of people initially categorized as 'multiracial' (based on two or more racial category selections) by about 50%, with greater portions of 'multiracial' individuals selecting racialized minority groups as their 'best represents' race as opposed to White (e.g., although the NH White population in the 'select all that apply' race variable is about 75%, less than 40% of individuals select White as their 'best represents' race)" (Franco et al., 2022).

Fourth, the CDC has used statistical methods since the 1980s to adjust BRFSS data to reflect population proportions on age, race and ethnicity, and sex, in addition to other demographic variables (*CDC - About BRFSS*, 2019). However, sweeping categorizations of populations without any clear and valid evidence to support the relevance of racial social

construction in public health research is antithetical to fundamental scientific principles. Potential bias resulting from selection bias due to race cannot be accurately corrected through weighting when racial construction itself is a major source of bias. In addition, data based on race are not interpretable without a meaningful and applicable construct.

Table 6: Race Questions in Behavioral Risk Factor Surveillance System (BRFSS)	
	Question and Response Options
US states, the District of Columbia, and three U.S. territories	"Which one or more of the following would you say is your race?" White Black or African American American Indian or Alaska Native Asian (*options for*: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian) Pacific Islander (o*ptions for*: Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander) Other* No choices* Don't know/Not sure* Refused*
If Multiple Races Selected	"Which one of these groups would you say best represents your race?" *Options same as above*
Reactions to Race Module	White Black or African American Hispanic or Latino Asian Native Hawaiian or Other Pacific Islander American Indian or Alaska Native Mixed Race* Some other group Don't know/Not sure* Refused*
**Survey administrator did not read. Appears as coding option.*	

8. Scale Development

This study involved the development of a novel tool for critical analysis – the Critical Race Framework. Boateng and colleagues developed a nine-step scale development framework that is intended for social, psychological, and health behaviors and experiences (Boateng et al., 2018). They defined three phases and nine steps of scale development: A) Item development: 1) domain and item generation, 2) content validity; B) Scale development: 3) pre-testing of

questions, 4) sampling and survey administration, 5) item reduction, 6) extraction of factor, C) Scale evaluation: 7) testing of dimensionality, 8) testing of reliability, and 9) testing of validity (Boateng et al., 2018). A major contribution of their publication was that it provided an accessible and structured approach. It is also a highly cited study on scale development, garnering 2,809 citations as of November 29, 2023.

There is no consistent scale for critical analysis tools. Some scales include yes-no options while others include qualitative assessment on a low-high scale (Eldridge et al., 2016; Leary et al., 2011; Munn et al., 2014; Pluye & Hong, 2014). We considered theoretical and practical factors through deductive (e.g., literature review) and inductive (e.g., expert feedback) methodologies for scale selection: 1) currently used in the literature, 2) whether scale options are interpretable, 3) respondents' ability to distinguish between options, 4) experiences in article critiques, 5) impact on survey response (e.g., survey fatigue), and 6) impact on analysis. A lengthy number of questions combined with considerable Likert scales (e.g., five to seven items) can lead to survey fatigue, which is associated with suboptimal survey quality and premature study termination (Lavrakas, 2008).

Boateng and colleagues' primer provided a stepwise approach for scale development in public health and social sciences. However, a major disadvantage appears in the linear nature of instrument development. Real-world conditions can require an iterative or flexible approach, particularly in the necessity of returning to previous steps or operating under limited resources. Their scale also has limitations for innovative instruments used in critical analysis since articles and studies, rather than human participants, form the basis for reliability and validity analysis. Generally, reliability and validity evidence for critical appraisal tools accrue over several years.

This study applied their framework, where appropriate, for development of the Critical Race Framework.

9. Summary of Chapter 2

Chapter 2 discussed critical perspectives on health research, including common misinterpretation of study findings. It also discussed the results of our literature review that illustrated the need for a bias tool since only one appears in the literature – the medical education field. None appears in the public health literature. The CARMeL Tool is innovative, but lacks a complete approach to applied scientific reasoning. It does not sufficiently explain the sociopolitical context to justify racial grouping. The tool largely serves as a continuation of current research practices wherein the racial paradigm is upheld under specific criteria. We explained how BFRSS data integrity in compromised due to its own race conceptualization, data-gathering, and analytical decisions. Chapter 3 discusses the dissertation study design to improve approaches to evaluating the quality of research.

Chapter 3: Methods

1. Introduction to Chapter 3

A major gap appears in the literature for assessing the quality of research due to threats that racial variables pose in key aspects of scientific quality (Lin & Kelsey, 2000; Witzig, 1996). A single tool appears in the medical literature (Garvey et al., 2022). No tool appears to exist in the public health literature for structured critical appraisal. This study developed the Critical Race Framework training and tool to address this gap.

A summary of the research questions, aims, methodologies, and designs by phase appears in Table 7. Phase I was a pilot study of the Critical Race Framework 1.0 that was conducted among faculty and doctoral students at a single site. Phase II was a national cross-sectional study of the Critical Race Framework 2.0 among public health experts. Phase III involved three public health researchers applying the Critical Race Framework 2.0 to twenty public health studies. We organized our subsequent discussion by phase and components within which we discussed methodology, design, study sample, recruitment, data collection, instruments, and analyses. Each component is discussed in the following order by phase: A) method, B) design, C) study sample, D) sample size, E) recruitment, F) procedures, G) data collection, H) instruments, I) analysis, and J) benefits and risks.

This research study was reviewed according to the University of Maryland, College Park Institutional Review Board (IRB) procedures (No. 2010403-1). The IRB determined that this study was educational and exempt from human subjects research. An amendment was filed in August 2023 to modify the study to add a compensation schema for Phase II participants and to remove focus groups that were initially planned for Phase III. In Phase II, individuals with full

responses were eligible for one of five random drawings for $100. The revised Phase III study design did not constitute human subjects research and did not require an IRB amendment.

Table 7: Summary of Research Questions, Aims, Methodologies, and Designs by Phase			
	Phase I	**Phase II**	**Phase III**
Research questions	**RQ1.** To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs? **RQ2.** What improvements to the Critical Race Framework 1.0 study design, instrument, and training are needed to improve?	**RQ1.**To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs? **RQ3.**What are demographic and research associations with perceptions of the Critical Race Framework? **RQ4.** What is the a) reliability and b) validity evidence for the Critical Race Framework?	**RQ4.a)** What is the reliability evidence for the Critical Race Framework? **RQ5.** What is the quality of 1) highly cited health disparities studies and 2) behavioral health studies based on the Critical Race Framework?
Aim	**A1.** To determine the extent to which the Critical Race Framework web-based training and bias tool are acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs. **A2.** To determine the improvements to the Critical Race Framework that are needed to improve?	**A1.** To determine the extent to which the Critical Race Framework web-based training and bias tool are acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs. **A3.** To determine whether demographic and research factors are associated with perceptions of the Critical Race Framework. **A4.** To determine the quality of reliability and validity evidence for the Critical Race Framework.	**A4.** To determine the quality of reliability evidence for the Critical Race Framework. **A5.** To determine the quality of highly cited health disparities studies and behavioral health studies based on the Critical Race Framework.
A. Method	Survey research Development Framework	Survey research Development Framework	Article critical appraisal
Design	Cross-sectional design	Cross-sectional design	Series of article analyses
Study Sample	Census of University of Maryland School of Public Health faculty and Department of Behavioral and Community Health doctoral students	Public health experts with terminal degree and high self-assessment on familiarity with scientific concepts	Public health researchers (non-human-subjects research) Sample of articles
Recruitment	Sampling via email and discussion forums	Sampling via email and discussion forums Snowball sampling	—
Eligibility Screening	Affiliated with the University of Maryland	High familiarity with research core concepts, terminal degree in public	—

	School of Public Health, familiarity with research core concepts and theory, terminal degree in public health or doctoral student, 18 years of age or older	health, 18 years of age or older, reside in the United States	
Data Collection	Qualtrics survey platform	Qualtrics survey platform	Google forms Word document
Instruments	Demographic and research experience CR Framework responses and evaluation of fit measures (acceptability feasibility, appropriateness, relevance, and satisfaction)	Demographic and research experience CR Framework responses and evaluation of *measures of fit* (acceptability feasibility, appropriateness, relevance, satisfaction, and target audience)	CR Framework responses
Analysis	Data distribution analysis Tests of independence Missingness analysis Non-differentiation analysis Qualitative analysis User data analytics	Testing of statistical assumptions Tests of independence Missingness analysis and imputation Measures of scale reliability Non-differentiation analysis Bivariate analyses (comparing means, correlation) Content validity analysis Construct validity analysis Exploratory factor analysis User data analytics Qualitative analysis	Testing of statistical assumptions Interrater agreement Interrater reliability

2. Research Questions

The research questions for this dissertation study were as follows:

RQ1. To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs? (Phase

I-II). *Aim 1*) To determine the extent to which the Critical Race Framework web-based training and bias tool are acceptable, satisfactory, feasible, appropriate, and relevant to *a priori* constructs.

RQ2. What improvements to the Critical Race Framework 1.0 study design, instrument, and training are needed to improve? (Phase I). *Aim 2*) To determine the needed improvements to the Critical Race Framework study design, instrument, and training?

RQ3.What are demographic and research associations with perceptions of the Critical Race Framework 2.0? (Phase II). *Aim 3*) To determine whether demographic and research factors are associated with perceptions of the Critical Race Framework.

RQ4. What is the reliability and validity evidence for the Critical Race Framework 2.0? (Phase II). *Aim 4*) To determine the quality of reliability and validity evidence for the Critical Race Framework.

RQ5. Article Analysis. What is the quality of 1) highly cited health disparities studies and 2) behavioral health studies based on the Critical Race Framework? (Phase III - Article Analysis). *Aim 5*) To determine the quality of highly cited health disparities studies and behavioral health studies based on the Critical Race Framework.

3. Pre-Study – Dissertation Phase

The section describes the methods for the development of the Critical Race Framework tool and training, consistent with our conceptual model (Steps 1-4). Steps 5-14 (Phases I-III) are covered in their respective methodological sections below.

A. Methods for Instrument Development.

Step 1. Identification of Domains – We identified four *a priori* themes for critical appraisal – reliability, validity, internal validity, and external validity. The basis for these constructs were discussed in Chapters 1 and 2.

Step 2: Item and Scale Generation 1.0 – The PI drafted the first iteration of the Critical Race Framework in December 2022. He reviewed the literature on critical evaluation to formulate critical questions. He then constructed questions to address the use, collection, and analysis of race variables in research in the four a priori areas. The deductive method was applied consistent with use of the literature to capture items for a domain (Boateng et al., 2018).

The initial CR Framework had 108 questions, which were highly detailed (not shown). The considerable list of questions is consistent with recommendations for tool development to be at least "twice as long as the desired final scale" (Boateng et al., 2018; Kline, 2013). However, this study had two delimitations in narrowing the scale that is not common in scale development: 1) a single PI within a dissertation study who made decisions to eliminate and modify items and 2) the absence of an expert panel to help guide these decisions. In published scale development models, the panel would have provided content validity at this stage of development (e.g. Delphi technique). Substantive expert engagement occurred in Phases I and II.

The initial CR Framework, referred to as *Item and Scale Generation 1.0* in the conceptual model, covered considerable domains and subdomains. For example, the historical threat to internal validity alone had eight questions: 1) "Evaluated differences in prior histories of racial groups that could influence variables of interest for single assessment," 2) "Evaluated differences in prior histories of racial groups that could influence changes in variables over time," 3) "Evaluated differences in prior histories of racial subgroups that could influence variables of interest," 4) "Evaluated prior exposure to racism as a historical threat," 5) "Evaluated differences

in prior histories of racial subgroups that could influence variables of interest for single assessment," 6) "Evaluated differences in prior histories of racial subgroups that could influence changes in variables over time," 7) "Discussed exposure to racism during study as a historical threat," and 8) "Discussed relevance of historical trauma as historical threat."

The initial scale for the CR Framework was three items ("no discussion," "low-quality discussion," and "high-quality discussion") in addition to a "not applicable" option. This was a modification of a rubric used for assessing library learning resources that used a five-item Likert scale on quality (low to high) (Leary et al., 2009, 2011; Yuan & Recker, 2015). A "no discussion" option was added because the literature supported that there was under-reporting about racial conceptualization and measurement (Martinez et al., 2022). However, there is no scientific consensus on the appropriate scale for critical analysis or bias tools for research reviews (Eldridge et al., 2016; Leary et al., 2011; Munn et al., 2014).

A "yes" response indicated that the article covered the topic at some point during the article. It could appear in any section – the background, methods, discussion, conclusion, imitations, or supplemental materials. A "no" response indicated that the article did not cover the topic from that question, meaning it did not appear anywhere in the article or supplemental materials. The choice of "unclear" meant that there was some ambiguity. "Unclear" was not to be confused with "did not discuss" or "did not assess". If the article did not discuss, assess, apply, interpret, or report certain information, then it should have been rated as a "no" response. On the hand, if the article suggested language that made it unclear and not a clear "yes" or "no," then "unclear" was appropriate. The final option is "not applicable". An article in which a question prompt does not pertain to the study might include a study that did not calculate p-values.

Step 3: Dissertation Committee Review – In December 2022, the dissertation committee shared concerns about the length of the 108-item CR Framework because it presented likely practical challenges during study implementation. The committee did not prescribe a specific length. Content validity was not evaluated at this stage of development.

Step 4: Item and Scale Generation 2.0 – The PI revised the CR Framework, as indicated by *Item and Scale Generation 2.0* in the conceptual model, in January 2023 to reduce the length considerably as per the dissertation committee's concerns. The new framework numbered 30 items. The number of total items was guided by the length of similar critical analytic tools or reporting checklists: the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-2020)(items =42), Joanna Briggs Institute Prevalence Critical Appraisal Tool (items = 10), and Meta-analysis Of Observational Studies in Epidemiology (MOOSE) reporting (items=35) (Munn et al., 2014; Page et al., 2021; Stroup et al., 2000). The CR Framework 2.0 numbered 30 items. The final scale comprised options for "yes," "no," "unclear," and "not applicable".

The PI retained the four critical areas. The reliability section included three questions while the validity section included six questions. The section on internal validity had 18 questions. They are intended to help identify threats to internal validity due the operationalization of theory, inherent racial measurement error, and the error arising from how race is handled in analysis. External validity covered population, temporary, and ecological validity considering racial heterogeneity and indeterminacy.

Methods for Theory-Informed Training Development. We applied Knowles' adult learning theory and Bryan's public health theory adoption to inform development of the web-based training on the CR Framework (Bryan et al., 2009; Knowles, 1980, 1984; Knowles et al.,

2014). Prior to explaining how to apply the Framework, the study discussed the importance of learning new skills or content from the research activity. It identified four relevant areas: spotting potential issues in research, refining their own research techniques ("personal payoff"), and appealing to personal motivation and lifelong learning. It also integrated principles that adult learners should be involved in the content development and evaluation by encouraging study participants to jot down their feedback and reflection throughout study participation, which supported ongoing improvement of the CR Framework and related training. The introduction also addressed Bryan and colleagues' guidance for adult learning in public health practice by involving learners as stakeholders through ongoing improvement and define a key aspect of public health practice to assess the quality of research quality (Bryan et al., 2009). Participants were asked to provide extensive quantitative and qualitative feedback on the Critical Race Framework training for improvement consistent with adult learning theory.

4. Phase I

Phase I addressed two research questions – **RQ1)** To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to research constructs and **RQ2)** What are the needed improvements to the Critical Race Framework study design, instrumentation, and training?

Phase I was a pilot study to test acceptability, satisfaction, feasibility, and appropriateness for Phase II implementation. Consistent with pilot studies, RQ1 sought to test the feasibility of implementing an innovation and the likelihood of challenges with subsequent research stages (In, 2017). Phase I occurred at the University of Maryland at College Park School of Public Health.

A. Method. This phase of the study used two methodological approaches related to: *Method 1)* conducting a survey and educational training and *Method 2)* refining the CR Framework.

Method 1: Cross-Sectional Study – Our educational survey research and training aimed to identify evidence of acceptability, feasibility, satisfaction, relevance, and appropriateness. This phase occurred between May through June 2023. It included four district parts – a) eligibility screening, b) a pre-survey, c) training on the Critical Race Framework 2.0 tool and applying the CR Framework to a published article, d) providing framework ratings and post-survey on measures of acceptability, appropriateness, feasibility, relevance, and feedback on the tool. The training was divided into modules for each of the four core concepts in addition to an introduction video. Most modules were under three minutes. The length of the training video was slightly under 15 minutes.

Method 2: Refinement of the CR Framework Study Design, Training and Tool

The second aim of Phase I was to improve the Critical Race Framework study design, training, and tool for Phase II. Results from Phase I were used to identify improvement in these areas. Our methodological approach was described in Chapter 2 and the *Pre-Dissertation* section of Chapter 3. We continue our discussion of the *Conceptual Framework for Critical Race Framework Development* (Steps 6-9) in Table 8.

We were interested in whether an article critique *study design* could be improved. We relied on pilot Phase I data results for this determination. In Phase II, we initially planned for each of a dozen researchers to perform 8-10 article critiques. Articles would have been drawn from the bibliographic analysis by Christopher Murray and colleagues (Arul & Mesfin, 2017).

The *instrument* concerned the CR Framework itself. This was a document with questions and a common rubric (Table 8). The *training* pertained to the web-based training video in Phase I. We sought data on the appropriate length, quality, and areas of improvement.

B. Design. The design of Phase I was a cross-sectional survey and article critique among public health faculty and doctoral students at a single site.

Table 8: Conceptual Framework for Critical Race Framework Development (Phase I – Steps 5-9)

- **Step 5. Content Validity** – Step 5 was a judgment-quantification step in which public health experts rated the relevance for each item in the CR Framework, then expert judgment was quantified using CVI. Content validity often appears in the literature as a criterion-based process (Larsson et al., 2015). However, this study did not intend to develop a critical appraisal tool to fully capture all subdomains of each construct. A fully developed instrument in this respect may adversely impact *measures of fit* and implementation effectiveness.

- **Step 6. Fit Evaluation of CR Framework** – Fit was determined by performance of six measures – (1) acceptability, (2) appropriateness, (3) feasibility, (4) satisfaction, (5) relevance, and (6) target audiences. Strong validity evidence appear in the literature for items 1-3 (IAM, 2020).

- **Step 7. Article Evaluation** – This step involved application of the CR Framework to a single research article.

- **Step 8. Single-Investigator CR Framework Refinement** – This stage involved synthesis of study results to inform three levels of modification: the study design, instrument, and training. Following merging of Parts 1 and 2 data sets, all data were cleaned and analyzed prior to initiating refinement.

    - **Study Design:** The initial study design for Phase II was a series of 8-10 article critiques for 6-10 co-researchers, in addition a pre- and postsurvey. We needed to determine the time that Phase I respondents spent on a single article critique to estimate the total time for the proposed Phase II study design. Quantitative data based on *measures of fit* and qualitative data were assessed to determine the practicality of this study design for Phase II.

    - **Training:** In revising the training video, we relied on and retained Knowles' adult learning theoretical work as a foundation. We did not identify a prior framework by which revision to the training portion was revised. Quantitative and qualitative study results informed strengths and weaknesses of video-based training in Phase I.

    - **Instrumentation** – We defined a process of refinement in five major areas: A) structure, B) scale, C) question retention/revision and deletion, D) addition of contextual cues, and E) length.

- **Step 9. Item and Scale Generation 3.0** – Following Steps 5-8, Item and Scale 3.0 of the Critical Race Framework is produced.

Note: These steps are not necessarily sequential, as they all occur during Phase I.

.

C. Study Sample. The Phase I sampling method involved a census of University of Maryland (UMD) School of Public Health (SPH) faculty and doctoral students in the UMD SPH Department of Behavioral and Community Health (BCH). This sampling frame was chosen because it was a convenient population given the location of the study. The study design and sample population are justified. Determining whether measures are theoretically linked is qualitatively judged by experts (Weiner et al., 2017).

D. Sample Size. The minimum sample size of five experts for content validity is consistent with the literature (Larsson et al., 2015; Lynn, 1986). We desired a sample size of 15 to gather sufficient feedback on the Critical Race Framework since this was the first occasion for experts to engage with the framework. We could not establish with confidence the expected enrollment rate and attrition to satisfy the minimum sample size of five experts.

E. Recruitment. This study relied on faculty emails listed on the University of Maryland School of Public Health website for each of the seven departments/units (Behavioral and Community Health, Epidemiology and Biostatistics, Family Science, Health Policy and Management, Kinesiology, Maryland Institute for Applied Environmental Health, and Public Health Science)(*Faculty & Staff | University of Maryland | School of Public Health*, n.d.). PhD students in the Department of Behavioral and Community Health (BCH) were also recruited. The BCH graduate program director forwarded the invitation via email to the doctoral students' listserv. The PI sent an individual invitation to approximately 280 faculty members in the UMD School of Public Health in addition to roughly two-dozen BCH PhD students. Individuals who appeared on the faculty website and had the following titles were excluded: emeritus, faculty specialists, program manager, business manager, faculty assistant, program advisor, coordinator, director of finance and administration, anything finance and administration, anything assistant,

academic advisor, assistant director, recruitment advisor, head coach, scheduler, facilities manager. The first wave of solicitations took place on May 2, 2023. We sent the second solicitation to non-enrollees on May 13, 2023. Following IRB determination, Phase I began study recruitment, which coincided with the end of the academic year.

F. Procedures. Participants were screened for eligibility: over 18 years of age, a faculty member in the University of Maryland School of Public Health (SPH) or a PhD student in the BCH Department and fluent in English. After screening for eligibility and providing informed consent, participants answered questions on demographics, background (number of year teaching, number of years since master's degree, name of graduate school), and research experience.

Participants then took part in the web-based asynchronous training by module – reliability, validity, internal validity, and external validity. The format of the web-based tutorial was an embedded video that plays when the user pressed the play button. A transcript of the audio was provided on the screen. At the end of each module, participants indicated one of the following: "I'm ready to continue", "I do not have any questions" or "I'm ready to continue. I have some questions (open comment)". They then received four prompts: "Approximately how long did it take to complete the Critical Race Framework training from start to finish, including reading the CR Framework and watching the videos"; "Please rate your satisfaction with the Critical Race Framework training"; "Please provide any feedback on the Critical Race Framework training"; "Provide email".

Participants received a web-based 15-minute tutorial on using the CR Framework, then were assigned an article to read - *Eight Americas: New Perspectives on U.S. Health Disparities* by Christopher Murray et al (Murray et al., 2005). This article was randomly generated from a

bibliographic analysis by Arul and Mesfin (Arul & Mesfin, 2017). Participants were asked to apply the CR Framework and to respond to a questionnaire to include measures of acceptability, appropriateness, and feasibility, as well as measures associated with clarity and for each primary domain. To "apply the CR Framework" means to complete the document labeled, "Critical Race Framework". Participants completed the CR Framework 1.0 and survey in Qualtrics for each article. Participants had two weeks to complete the training, to read the article, and to complete the online assessment. Participation was expected to take 45-90 minutes, depending on level of expertise – 15 minutes for web-based training, 15 minutes to complete article, and 15 minutes for highly efficient and skilled study participants. Prior to beginning article review, participants needed to have indicated that they did not have any questions. All Phase I activities – training, downloading article, completing CR Framework and survey questions – took place in Qualtrics. Participants were required to submit questions and receive answers prior to starting article review.

G. Data Collection. Survey data were collected and stored in Qualtrics. Any potential loss to confidentiality was minimized by keeping all data in a password-protected virtual storage location. The University of Maryland storage site (UMD BOX) was highly secure and HIPAA compliant. Participants' information may have been shared with representatives of the University of Maryland or governmental authorities if we observed that participants are in a life-threatening danger, or if we were required to do so by legal statutes. Only the principal investigator had access to the study data. Data from ineligible responses were destroyed before analysis in 2024 and not used for research purposes.

H. Instruments. This study assessed researcher characteristics and perceptions, including acceptability, appropriateness, and feasibility of the Critical Race Framework. This list does not include the exclusion criteria.

1. **Gender –** The following options for gender could be selected: man, woman, two-spirit, non-binary, I prefer not to answer, and none of the above. Man, woman, and non-binary appeared randomly for each participant.
2. **Social Identity –** Participants were asked to respond to a question on social identity, "What are your most important social identities? Social identity is a person's sense of who they are based on their group membership(s)" (Tajfel & Turner, 1979).
3. **Acceptability** - Acceptability is agreement among study participants whether the Critical Race Framework is agreeable or satisfactory (Proctor et al., 2011). We used the Acceptability of Intervention Measure (Cronbach's $\alpha \approx 0.9$) (IAM, 2020).
4. **Graduate Institution** - Name of school from which respondent obtained their master's degree.
5. **Completion Time** – Four separate questions asked about 1) time to complete Part 1 of Phase I, 2) time to read the Critical Race Framework, 3) read the assigned article from Phase I, and 4) complete the CR Framework for the article.
6. **Research** – Number of years having conducted research.
7. **Race in Research** – Whether respondent has ever used a variable for race.
8. **Assumption about Race** – Optional open-text field to discuss personal implicit or explicit assumptions that were race used in research.

9. **Appropriateness** – The CR Framework's fit for practitioners, relevance to practice, and compatibility (Proctor et al., 2011; Weiner et al., 2017). We used the Intervention Appropriateness Measure (Cronbach's $\alpha \approx 0.9$) (IAM, 2020).
10. **Feasibility** - The extent to which the Critical Race Framework is practical and viable (Proctor et al., 2011). We used the Feasibility of Intervention Measure (Cronbach's $\alpha \approx 0.9$) (IAM, 2020).
11. **Satisfaction** - Single-item Likert scale for the web-based training and tool, "Please rate your satisfaction with the Critical Race Framework training". The options were strongly disagree, disagree, neutral, agree, strongly agree.
12. **Familiarity** with each core construct (i.e., reliability, validity, internal validity, and external validity) was also used in analysis. The five-item scale ranged from not familiar to very familiar.
13. **Quality of Evidence Scale –** Four questions about quality of evidence in reliability, validity, internal validity, and external validity using four-item scale (very low, low, moderate, and high).
14. **Completing Critical Race Framework –** Provided responses to Critical Race Framework based on assigned article.
15. **Relevance** is assessed by asking how relevant each question in the Critical Race Framework is to its related construct.
16. **Quality of Question –** Asked to indicate if a question of the CR Framework was associated with any of the following: "I had difficulty understanding this question," "This question should be reworded for clarity," and "This question is redundant."
17. **Qualitative Responses –** We provide four qualitative data collection forms.

- "Please discuss what implicit or explicit assumptions that you made about race in your research." (Part 1)
- "I'm ready to continue. I have some questions. Someone from the study will provide responses within 2 business days. Please do not complete CR Framework for the article analysis before you hear back from us." This appeared after each model of the training." (Part 1)
- "Please provide any feedback on the Critical Race Framework training." (Part 1)
- "Please provide any additional feedback on the Critical Race Framework." (Part II)

I. Analysis.

A) Data Cleaning, Checking of Assumptions, and Imputation

i. Data Cleaning. Prior to exporting the SPSS file from Qualtrics, the data were recoded. The scales were verified after importing and before excluding responses from analytic sample. Duplicate study participants based on name and email were removed. In such instances, we retained respondent's most complete response. We excluded respondents that did not initiate any activities after study enrollment. Individuals who indicated that they were ineligible for the study based on any survey response after eligibility screening were excluded.

ii. Checking Assumptions. We conducted tests for homogeneity of variance, equality of variance between group normality, presence of outliers, kurtosis, and skewness to determine the appropriate test and analysis.

iii. Missingness. Studies involving missing data are common and can be prone to inaccurate findings without adequate attention to missing values (Graham, 2009). We relied on the decision tree by Mirzaei and colleagues for handling missing data (Mirzaei et al., 2022). We

computed percentage of missingness at the item-level as opposed to the scale-level (e.g., scale summation score) since imputation at the scale-level tends to reduce precision and may require a significant increase in sample size (Mazza et al., 2015).

There are three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Multiple imputation can be used for MCAR and MAR (Bhaskaran & Smeeth, 2014). Multiple imputation can be considered inappropriate for MNAR (Bhaskaran & Smeeth, 2014). However, imputation is viable with MNAR under certain conditions. Gomer concluded, "Our findings indicate that most methods for handling missing data can tolerate a certain proportion of missing values that are not from the assumed missing data mechanism. For example, in a regression model with three predictors and 31% missingness in a predictor variable, ML and MI can tolerate up to 20% of missing values that are actually MNAR. Using MI or ML, the bias of adjusted R-squared and the inflation of Type I error rates can be minor. Although MNAR values are never desirable, our findings suggest that they are not harmful in small amounts" (Gomer, 2019). However, we found no evidence to justify multiple imputation for missingness beyond 50%.

Conducting analysis for MNAR data sets is less straightforward, "As the data that is missing depends on the missingness itself, it becomes difficult to use any other information to impute or infer the value. Most imputation methods result in biased results when dealing with MNAR. Suggestions for handling MNAR data are to utilize (sic) questions in order to find a related factor, assess the theoretical consideration for using the question or variable, or consider dropping the variables when performing analysis" (Mirzaei et al., 2022). MNAR is non-ignorable missingness, meaning that there could be "an unobservable factor that is directly affecting the reason that the data values are missing. This can be the question itself being the

cause of the missing response, or underlying assumptions" (Mirzaei et al., 2022). In complete case analysis, MNAR data may or may not bias results, but this scenario is likely to be biased (Mack et al., 2018). "Complete case analysis will be unbiased due to missing data if the missingness is independent of the outcome under study, a condition that can be present whether the data are MAR or MNAR" (Mack et al., 2018). In the context of this study, handling missing data that are MNAR relied on subjective analyses, in consultation with dissertation committee members.

For missingness less than ≤5%, we intended to conduct Little's Missing Completely at Random (MCAR) Test to determine whether imputation was appropriate. It is impossible to test the MAR assumption that missingness is associated with another variable in the data set, "It is a fundamentally untestable assumption, because it concerns the unobserved values" (Bhaskaran & Smeeth, 2014). A determination of MAR is based on subjective analysis. Bhaskaran and colleagues illustrated the conundrum using this example, "(Y)ou would need to know the distribution of blood pressure among people with no blood pressure recorded, to compare it with the distribution of blood pressure among those with complete data. So it really is an assumption that you need to justify based on background knowledge and discussion with experts" (Bhaskaran & Smeeth, 2014). In the context of this study, data that are MAR would be plausible if, for example, men were less likely to complete the CR Framework portion of the study and that perceptions of the CR Framework did not otherwise differ between men and women.

Mirzaei and colleagues provided an inconclusive finding for handling MNAR data, "Suggestions for handling MNAR data are to utilise *(sic)* questions in order to find a related factor, assess the theoretical consideration for using the question or variable, or consider dropping the variables when performing analysis" (Mirzaei et al., 2022). Several steps were

required before defining the analytic sample. First, data sets from Parts 1 and 2 were combined to compute the missingness percentage. Variables for missingness analysis were identified. We determined whether the data were MCAR, MAR, or MNAR. We used complete case analysis for MNAR data, as discussed.

B) Potential Threats to Reliability

We discussed the scope of our theoretical model for reliability in Chapter 1 (Table 3). Satisficing behavior is a decision-making strategy in which respondents make an adequate or satisfactory rather than an optional choice. Non-differentiation or straight-lining is a common indicator of satisficing behavior. Straight-lining is defined as a tendency to select an identical response for items in a set of questions (Kim et al., 2019). There is no consensus on a standard for measuring straight-lining behavior (Kim et al., 2019). Due to multiple scales and repeated questions on relevance used in this study, participants may be prone to satisficing behavior. To analyze straight-lining, we calculated the simple nondifferentiation measure by computing the percentage of study respondents who used an identical response at the scale-level for *measures of fit* and construct-level on relevance and evaluation responses. Higher nondifferentiation may indicate more straight-lining (Kim et al., 2019).

Ambiguity in the Critical Race Framework questions or terms can cause systematic errors in survey data. To assess potential response bias arising from inaccurate responses, we asked participants to indicate if any question in the Critical Race Framework satisfied any of the following conditions: “I had difficulty understanding this question”, “This question should be reworded for clarity”, or “This question is redundant.”

C) Data Analysis

We provided a descriptives table for data variables. Frequency tables for satisfaction, acceptability, appropriateness, and feasibility were generated using complete case analysis. A table describes duration for each component of Phase I and by familiarity and role. We defined sixty percent (60%) or more of respondents to determine whether the Critical Race Framework was acceptable, feasible, appropriate, and satisfactory, assuming an adequate sample size. We interpreted findings on these measures considering potential threats to reliability (e.g., nondifferentiation). The qualitative data were classified as pertaining to a priori themes: study design, instrumentation, or training.

J. Benefits and Risks. There were direct benefits from participation in this research. New critical analysis skills for evaluating how race conceptualization, collection, and analysis were potential benefits. The CR Framework itself may have been used for personal research practices and analysis. The information that participants provided may help researchers better understand informed approaches for assessing the quality of research.

There was minimal risk from participating in this research study. Participants may have felt discomfort or embarrassment answering questions about a rubric about race in research. The investigators mitigated the risk of discomfort or embarrassment by allowing participants to skip some questions. There was also the risk of breach of confidentiality. This risk was mitigated by relying on secure, password-protected systems at the University of Maryland- College Park.

5. Phase II

A. Method. Phase II was a national cross-sectional study to collect survey data on perceptions of the Critical Race Framework 2.0 and participants' public health background. We conducted this study to establish evidence of reliability, validity, and fit ("*measures of fit*").

Participants completed a one-time web-based training on interpreting the CR Framework and responded to a questionnaire to include measures of acceptability, appropriateness, and feasibility, as well as measures associated with clarity and for each primary domain. Demographic data and research experience were collected prior to the training. Phase II methodology consisted of three major components: 1) eligibility screening, 2) survey, 3) web-based training, and 4) post-survey. The total time for Phase II was expected to be between 30-45 minutes. Table 9 discusses Steps 10-12 for the CR Framework conceptual model: content and construct validity, tests of reliability, and fit evaluation.

B. Design. Phase I was a national cross-sectional survey study.

Table 9: Conceptual Framework for Critical Race Framework Development (Phase II – Steps 10-12)

- **Step 10. Content and Construct Validity** – Content validity was assessed using relevance of each item in the Critical Race Framework. Although similar questions appeared in Phase 1, content validity was not a primary aim during this phase. Phase II was a judgment-quantification step in which public health experts rate the relevance for each item in the CR Framework, then expert judgment is quantified using CVI. We assess the construct validity for our *measures of fit*, which had shown strong validity in the literature (Proctor et al., 2011).

- **Step 11. Tests of Reliability** – Reliability theory concerns the accuracy and replicability of study results. This phase evaluated potential bias due to satisficing behavior, survey or instrument bias, and missingness analysis.

- **Step 12. Fit Evaluation of CR Framework** – We analyze data results on our six *measures of fit*, consistent with our conceptual model, to determine the overall fit of the Critical Race Framework.

C. Study Sample. Phase II was a national cross-section of public health experts who were: 1) 18 years of age or older, 2) were fluent in English, 3) residents in the US, and 4) at least "fairly familiar" with the concepts of reliability, validity, internal validity, external validity, and statistical theory and analysis. Participants needed a terminal degree in a public health specialty for eligibility. The study design and sample population are justified because determining whether

measures are theoretically linked is usually based on qualitative evaluation among experts (Weiner et al., 2017).

D. Sample Size. We conducted multiple analyses that each vary in determining the minimum sample size. The minimum sample size of five experts for content validity is consistent with the literature (Larsson et al., 2015; Lynn, 1986). The sample size for Cronbach's alpha coefficients in reliability analysis "can be as small as four with assumption of a very high internal consistency such as 0.95 or more" (Bujang et al., 2018). Published Cronbach's alpha coefficients for three *measures of fit* (i.e., acceptability, feasibility, and appropriateness) are high (Proctor et al., 2011).

Exploratory factor analysis (EFA) likely required the highest minimum sample size and set the overall minimum for this study. As a rule of thumb, EFA requires an absolute minimum of 10 observations per variable. Some absolute minimum sample sizes have also been proposed (Mundfrom et al., 2005). Taking the EFA (*EFA 3 in analysis*) with the highest number of variables (internal validity with 8 variables), it would mean that we needed 80 samples in our study for EFA 3. However, we discuss later the participant-to-variable ratio within the context of our study. This may not necessarily be the minimum sample size for reliably interpreting EFA results. An inadequate sample size for EFA would not invalidate other study results.

E. Recruitment. Several sampling strategies were used (Table 10). Participants with partial responses received 2-3 reminders to complete study. Those who had fully completed their survey were asked in a follow-up email to consider forwarding it to their network.

We posted to several forums to solicit participation and sought permission from forum moderators prior to posting – a best practice – unless no contact information was available and similar posts appeared in the history of posts. The study solicited via US-based public health

discussion forums, including but not limited to the American Public Health Association, Spirit of 1848, and Reddit (Table 10).

Second, we targeted contact authors from a bibliographic analysis of the most cited health disparities articles (Arul & Mesfin, 2017). We sent an email through an account set up for the study to the contact authors for all articles in the bibliographic analysis of Arul and Mesfin (Arul & Mesfin, 2017). We relied on the emails that appeared in the published article to send them an invitation from Qualtrics. Third, the dissertation committee members supported recruitment by forwarding the study solicitation to colleagues whom they believed may be interested in participating in the study. They forwarded the solicitation without modification. We described additional sampling strategies in Table 10.

Table 10: Summary of Recruitment Strategies (Phase II)

Platform	Description	Date(s) for Solicitation
Group Email Listserv/Online Groups/Social Media		
American Public Health Association	We posted two survey solicitations on the APHA Member Central forum (>24,000 members).	September – October 2023
Spirit of 1848	An official caucus of APHA. Several survey solicitations were posted.	August – October 2023
Reddit	We posted solicitation requests in two online forums: biostatistics (>10,000 members) and public health (>60,000 members). Two solicitation requests were posted.	August – September 2023
ForbesBLK Forum	An online community for Black professionals. We posted in online forums for users interested in health using the mental-health and healthcare hashtags. Two reminders were posted.	September 2023
CR Framework X	We created a website and X (formerly Twitter) account for this study. Solicitation requests were available on these sites.	August – December 2023
Individual Invitation		
Health disparities researchers	Nearly 100 authors were identified through the bibliographic analysis by Aru and Mesfin (Arul & Mesfin, 2017). They received two solicitation requests and were asked to forward to their network.	August – October 2023
Departments and Schools of Public Health	We contacted individuals from schools or departments of public health. In most cases, these were deans. They received two solicitation requests and asked to forward to their network.	August – October 2023
Dissertation Committee's Network	Committee members forwarded the invitation to their professional networks.	September – December 2023
American Public Health Association Annual Meeting	Dr. Craig Fryer, dissertation committee chair, recruited in-person at the APHA Annual Meeting. He shared a flyer containing the QR code for the survey.	October 2023

We relied on snowball sampling as a recruitment strategy, which is associated with improved trust in the researcher and study (Jeanfreau & Jack Jr, 2010). Partial and full survey respondents were asked to forward to their network. School and department contacts were asked to forward to faculty who were likely interested in the topic. The dissertation committee helped to recruit for this study through their professional networks. We did not seek to recruit from within the University of Maryland for Phase II.

F. Procedures. Phase II consisted of three major components: 1) eligibility screening, 2) survey, 3) web-based training, and 4) post-survey. Participants were screened for eligibility: 1) 18 years of age or older, 2) held a doctoral degree related to one of the following fields: biostatistics, epidemiology, health promotion and education, health administration, environmental health, global health, maternal and child health, nutrition, health policy and management, behavioral science and health education, public health policy and management, community health, and international public health, 3) fluent in English, 4) currently resided in the United States, 5) at least "fairly familiar" with each of the following constructs: reliability, validity, internal validity, external validity, and 6) statistical theory and analysis.

Participants were provided with informed consent. The survey collected a secondary email contact, doctoral degree, years held doctoral and master's degree, employer type, employee role, and state of primary residence. Participants rated quantitative and qualitative research skills, number of years conducting public health research and teaching public health, public health specialty, number of published studies, typical number of racial categories used in research, frequency of research activities, agreement with the research premise, and influence of U.S. Office of Personnel Management (OPM) definitions on research practices. Two questions on gender and social identity were asked.

Participants were provided with a nine-minute training course on the Critical Race Framework 2.0. A transcript appeared on the same screen. Following the training, participants were asked to enter a code that appeared at the end of the video. Questions on satisfaction, appropriateness for different learners and settings, acceptability, appropriateness, feasibility, and relevance. A screen appeared explaining the next steps for the drawing. Participants with partial responses received emails encouraging completion. Respondents who completed the questionnaire were asked to forward it to their network.

Participants who completed all components of Aim 2 (Phase II) – training and survey – were entered into a drawing to receive one of five $100 gift card compensation. Within six weeks after the end of Phase II, participants were notified via email that they had been selected and to complete the address form on the Qualtrics website. Once the selected recipient of the drawing replied to the email, the PI emailed each recipient individually with the following message, “Thank you for participating in the Critical Race Framework Study. “Your name was randomly drawn to receive a $100 gift card compensation. In 7-14 business days, you will receive an email from the business manager at the University of Maryland School of Public Health Department of Behavioral & Community Health for additional information on receiving your gift card. If you have any questions, please contact that department at (301) 405-3727.”

G. Data Collection. Survey data were collected and stored in Qualtrics. Any potential loss to confidentiality was minimized by keeping all data in a password-protected virtual storage location. The University of Maryland storage site (UMD BOX) is highly secure and HIPAA compliant. Participants’ information may have been shared with representatives of the University of Maryland or governmental authorities if we observed that participants are in a life-threatening danger, or if we were required to do so by legal statutes. Only the principal investigator (CW)

had access to the study data. Data from ineligible responses were destroyed before analysis in 2024 and not used for research purposes.

H. Instruments. Phase II deployed a questionnaire, including validated instruments. These questions sought data on participants professional background and perceptions of the Critical Race Framework. The Phase II survey is available in supplement.

1. **Doctoral Degree** – Participants were asked to select from among 11 predefined degree types (e.g., Ed.D., Dr.P.H., M.D., D.O., D.Sc., D.P.E., D.B.H., D.H.Sc., D.S.W.) and an open-response "other" option. We dichotomized this variable (Ph.D. versus all other degree types) during data analysis.
2. **Years Held a Doctoral Degree** – Participants indicated a range of years for the time that they have held a doctoral degree (*coding values in parentheses*): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12). We treated this variable as continuous in analysis.
3. **Years Held a Master's Degree** – Participants indicated a range of years for the time that they have held a master's degree (*coding values in parentheses*): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12).. We did not use this variable in analysis in this stage of analysis due to the small sample size.
4. **Employer Type** – Several choices for employer type appeared: college/university, hospital/health system, nonprofit (other than higher education), for-profit, small business, self-employed, unemployed, and an "other" option (open response). This variable was dichotomized (college/university versus all other responses) for analysis.

5. **Government Employee** – A question was asked on government employment status (yes/no). We did not use this variable in analysis.
6. **Primary Role -** Primary roles included teaching faculty, non-teaching faculty, researcher (outside of academia), non-research administrator (outside of academia), and an option for other (open response). We dichotomized this variable in analysis for teaching faculty versus all other responses.
7. **US State/Territory** – A list of US states and territories for the participant's primary residence appeared. In analysis, states was grouped into one of four US regions based on the US Census Bureau Regions map (*US Census Bureau Regions and Divisions with State FIPS Codes*, n.d.). Puerto Rico was randomly assigned to region 1 (Northeast) using the Google random number generator since the US Census Bureau does not assign Puerto Rico to a region.

Research Experience

8. **Familiarity** with each core construct (i.e., reliability, validity, internal validity, and external validity) was also used in analysis. Each question was retained for analysis. A *Familiarity Index* consisting of five-item composite score was created using each self-assessed score on familiarity for each area of critical appraisal and familiarity with statistical theory and analysis. The maximum score is 20.
9. **Self-Assessed Qualitative and Quantitative Research Skills** – This question asked, "Please rate your quantitative and qualitive research skills". This question relied on a Likert scale: very poor, poor, average, very good, excellent with separate responses for qualitative and quantitative. Individuals' self-reported ratings were used to create the *Skills Composite Index*.

10. **Years in Public Health Research** - Participants indicated a range of years for the time that they have conducted public health research (*coding values in parentheses*): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12).. This was treated as a continuous variable in analysis.
11. **Years Taught in Public Health** - Participants indicated a range of years for the time that they have taught public health (*coding values in parentheses*): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12).. This was treated as a continuous variable in analysis.
12. **Number of Publications** – This required manual entry. A follow-up question asked respondents to confirm manual entry. This was treated as a continuous variable in analysis.
13. **Public Health Specialty** – While the eligibility screening asked if participants were from a list of public health specialties, this question asked them to indicate that specialty from a predefined list: behavioral science and health education, biostatistics, community health, environmental health, epidemiology, global health, health administration, health policy and management, health promotion and education, international public health, maternal and child health, nutrition, and public health policy and management. This listed was generated from a Perplexity search for public health specialties (*Perplexity*, n.d.). Two variables were created for analysis: 1) the number of specialties and epidemiology and 2) a binary variable whether a respondent indicated biostatistics or epidemiology versus all other specialties.

14. **Studies that Used a Variable for Race** – A range of the participants' percentage of published studies that included a variable for race included: none, less than 11%, 11-25%, 26-50%, 51-75%, 76-90%, >90%. This variable was dichotomized for 51% or more versus all other responses.
15. **Typical Categories for Race** – Participants provided the typical number of racial categories used in their research (i.e., less than 5, 5, 6, 7, 8 or more). This variable is not used in analysis.
16. **Racial Categories –** Participants has an option of naming racial categories other than the following that they used in their research: (1) American Indian or Alaska Native, (2) Asian, (3) Black or African American, (4) Native Hawaiian or Other Pacific Islander, and (5) White. This variable was not used in analysis.
17. **Research Practices –** Participants were asked to indicate frequency (e.g., never, rarely, sometimes, often, always) for conducting the following research practices: provide a definition of race to study participants, collapse racial groups in analysis, assume that race is stable, and consider potential threats to statistical inferences due to inclusion of a variable for race. This was treated as a continuous variable in analysis.
18. **Study Premise –** Participants were asked to indicate their agreement with the research premise, "The premise for this study is that race introduces potential threats to study quality in the areas of reliability, validity, internal validity, and external validity." We collapsed this variable: disagree ("disagree and "strongly disagree"), neutral, and agree ("agree" and "strongly agree").

19. **Influence of Federal Definition of Race –** Participants were asked to indicate the influence of the federal government's definition of racial categories. This was treated as a continuous variable.
20. **Perception of Federal Definition of Race –** Participants were asked to indicate their opinion of the federal definition of race: very negative, negative, neutral, positive, and very positive. This variable was collapsed: negative ("very negative opinion" and "negative opinion"), neutral, and positive ("very positive opinion" and positive opinion").
21. **Gender –** The following options for gender appeared: man, woman, non-binary, I prefer not to answer, and none of the above. Man, woman, and non-binary appeared randomly for each participant. This variable was collapsed into three categories: woman, man, and non-binary (including other responses).
22. **Social Identity –** Participants were asked to respond to a question on social identity, "What are your most important social identities? Social identity is a person's sense of who they are based on their group membership(s)" (Tajfel & Turner, 1979). We created a new variable based upon the number of social identities.
23. **Meeting Research Goal –** Participants indicated whether the Critical Race Framework training and tool met its research goal, "The goal of the Critical Race Framework is to develop a bias tool and structured approach to critically appraise a given health study that involves a race variable. Please rate how you feel the CR Framework training and tool met this goal." The scale was as follows: "did not meet this goal," "did not fully but mostly met this goal," "met this goal," and "exceeded this goal". This variable was collapsed: did not meet goal ("did not meet this goal" and "did not fully but mostly met this goal") and met goal ("met this goal" and "exceeded this goal").

24. **Appropriateness for Learner and Setting Type –** A list of learners, settings, and role types were shown for participants to indicate the appropriateness of the Critical Race Framework: Learners and role types included undergraduates, master's students, doctoral students, postdoctoral fellows, public health faculty, and researchers. Settings included classroom instruction, faculty development, journals, professional development, consulting, policymaking, public health planning, and evaluating funding requests. A summation variable was created for analysis.
25. **Video-Viewing Indicator –** The end of the Critical Race Framework video had a code that survey respondents were instructed to provide on a following screen. This is a likely indication whether they viewed the training video. We dichotomized those with correct codes versus no code or an incorrect code.

Measures of Fit

26. **Acceptability** is agreement among implementation stakeholders that a practice or treatment is agreeable or satisfactory (Proctor et al., 2011). We adopted the Acceptability of Intervention Measure (Cronbach's alpha ≈ 0.9) (IAM, 2020). A collapsed scale was retained for analysis (completely disagree/disagree, neither agree nor disagree, completely agree/agree, in addition to a second variable was created that summed the items for analysis). A maximum score of 25 indicated the highest agreement.
27. **Appropriateness** concerns fit for practitioners, relevance to practice, and compatibility (Proctor et al., 2011; Weiner et al., 2017). We adopted the Intervention Appropriateness Measure (Cronbach's $\alpha \approx 0.9$) (IAM, 2020). A collapsed scale was retained for analysis (completely disagree/disagree, neither agree nor disagree, completely agree/agree, in

addition to a second variable was created that summed the items for analysis). A maximum score of 25 indicated the highest agreement.

28. **Feasibility** is the extent to which a new practice or innovation is practice and viable (Proctor et al., 2011). We adopted the Feasibility of Intervention Measure (Cronbach's $\alpha \approx 0.9$) (IAM, 2020). The original scale was retained for analysis. A collapsed scale was retained for analysis (completely disagree/disagree, neither agree nor disagree, completely agree/agree, in addition to a second variable was created that summed the items for analysis). A maximum score of 25 indicated the highest agreement.
18. **Satisfaction** uses a single-item Likert scale for two questions on satisfaction, "Please rate your satisfaction with the Critical Race Framework training" and "Please rate your satisfaction with the Critical Race Framework tool". The options were strongly disagree, disagree, neutral, agree, strongly agree. Collapsed data were reported (strongly disagree/disagree, neutral, agree/strongly agree) A summation index using these two questions was created for analysis.
19. **Relevance** is assessed by asking how relevant each question in the Critical Race Framework is to its related construct. Each question of the Critical Race Framework is subsumed under one of four core concepts: reliability, validity, internal validity, and external validity. We dichotomized this variable for most analyses: low relevance ("not at all relevant" and "slightly relevant") and moderate-to-high relevance ("fairly relevant" and "highly relevant).
20. **Target Audience and Setting** determined for which audiences (undergraduates, master's students, doctoral students, postdoctoral fellows, public health faculty, public health researchers) and settings (classroom/instruction, faculty development, peer-reviewed

journals, professional development, consulting, policymaking, public health planning and implementation, and evaluating funding requests) for which the CR Framework is appropriate. This appeared as a checklist.

Qualitative Responses – We provided four qualitative response fields.

21. "What race categories did you use in your race data collection tool besides the following: (1) American Indian or Alaska Native, (2) Asian, (3) Black or African American, (4) Native Hawaiian or Other Pacific Islander, and (5) White."
22. "Please discuss what assumptions you typically made about race in your research (highly encouraged but optional response)."
23. Please use this space to reflect. (optional) If you have no reflections at this time, type "none" to proceed.
24. "You have completed all of the survey questions. You may provide any feedback on the Critical Race Framework training."

I. Analysis. This section describes the approach to data analysis – A) data cleaning, checking of assumptions, and imputation, B) descriptive statistics, and C) analyses. All analyses were performed in IBM SPSS version 29.0.

A. Data Cleaning, Checking of Assumptions, and Imputation

i. Data Cleaning. Prior to exporting the SPSS file from Qualtrics, the data were recoded. The scales were verified after importing and before excluding responses from analytic sample. Duplicate study participants based on name or email were removed. In such instances, we retained respondent's most complete response. We excluded respondents that did not initiate any activities after study enrollment. Individuals who indicated that they were ineligible for the study based on any survey response after eligibility screening were excluded.

ii. Checking Assumptions. We conducted tests for homogeneity of variance, equality of variance between group normality, presence of outliers, kurtosis, and skewness.

iii. Missingness. For missingness less than ≤5%, we conducted Little's Missing Completely at Random (MCAR) Test to determine whether imputation was appropriate. We imputed missing values or deleted listwise. For MNAR data, we did not impute and conducted complete case analysis.

B. Descriptive Statistics

We reported descriptive statistics overall and by partial and full respondents: participation rate, educational attainment, current role, familiarity with core constructs, number of years in public health research, gender, and social identities. Frequency tables for satisfaction, acceptability, appropriateness, and feasibility were generated.

C. Analyses

1. Bivariate Analyses.

Bivariate analyses were intended to answer RQ3. We compared groups using the Mann-Whitney *U* test for dichotomized variables based on demographic and research experience data. We conducted correlation analysis using Kendall's tau-b (τ_b) correlation coefficients for continuous or ordinal data.

2. Qualitative Analysis.

We thematically classified qualitative data. We had three *a priori* themes: study design, instrumentation, or training. The survey and article critique were the *study design*. The *instrumentation* concerned the CR Framework itself. The *training* pertained to the web-based training video in Phase I. Additional themes were identified during analysis. These data were used to identify evidence that would strengthen or undermine data results on *measures of fit*.

Open-ended questions included, “Please use this space to reflect (after training). (optional) If you have no reflections at this time, type "none" to proceed” and “You have completed all of the survey questions. You may provide any feedback on the Critical Race Framework training. Otherwise to complete the study and find about next steps, please click the bottom right arrow.”

3. Reliability Analysis.

We analyze straight-lining by calculating the simple nondifferentiation measure at the scale-level for *measures of fit* and construct-level for the four domains of the CR Framework — percentage of study respondents who used an identical response. Higher nondifferentiation may indicate more straight-lining (Kim et al., 2019).

We assessed measures of internal consistency for scale reliability analysis and reported Cronbach’s α for multi-item scales acceptability, feasibility, and appropriateness (IAM, 2020). We conducted tests comparing a consistency intraclass correlation coefficient (ICC) to absolute agreement ICC. Where the presence of non-negligible bias is likely, we reported multiple coefficients consistent with best practices (Bobak et al., 2018). We define good acceptability between 0.75 and 0.9 and excellent acceptability over 0.90 (Bobak et al., 2018).

4. Validity Analysis.

Content Validity Index. We computed a content validity index (CVI) for each item (I-CVI) and each construct (S-CVI). I-CVI and S-CVI the proportions of agreement on moderate to high relevance at the item- and scale-level, respectively (Lynn, 1986; Zamanzadeh et al., 2015). We divided the number of experts with acceptable scores (“fairly relevant” or “highly relevant”) by the total number of scores among raters. In CVI analysis by question (I-CVI), the denominator was the total number of raters. The denominator by construct was the number of raters multiplied by the number of questions for that construct (S-CVI). We consider an excellent

CVI to be ≥.75 consistent with cutoffs previously published in the literature (Larsson et al., 2015).

We calculated a modified Kappa value, using the formula: (k* = (CVI- Pc)/(1- Pc) (Rodrigues et al., 2017). The probability of chance occurrence (Pc) formula for a binominal random variable was used: Pc = [N!/A! (N–A)!]*.5^N (Rodrigues et al., 2017). *N* was the number of raters. *A* was the number agreeing on fairly or highly relevant. We used criteria proposed by Cicchetti and Sparrow to evaluate the Kappa values: Fair = k* of 0.40 to 0.59; Good = k* of 0.60 to 0.74; and Excellent = k* of 0.75 to 1.00 (Cicchetti & Sparrow, 1981).

Exploratory Factor Analysis. To analyze validity of the underlying constructs for the Critical Race Framework, we conducted exploratory factor analysis (EFA). As discussed, the Critical Race Framework is supported by four core concepts in research quality and critical appraisal. EFA has been used in validity and reliability analysis for newly implemented tools (Fitriana et al., 2022). Our analysis has several limitations due to the small sample size, MNAR data, and no clear standard for determining the minimum sample for our study type, as discussed below and in the Phase II results in Chapter 3.

There is no general agreement on the threshold for an adequate sample size in use of EFA (Beavers et al., 2019; Mundfrom et al., 2005; White, 2022). Some researchers favor large sample sizes, with an absolute minimum of 50 (de Winter et al., 2009). As a general heuristic, an absolute minimum of 10 observations per variable is required.

Conversely, some scholars have concluded, “Every step of the process in a factor analysis requires the researcher to be firmly grounded in contextual theory and fundamental understanding of factor analysis methodology. The greatest difference in methods centers on how the technique accounts for variance and relationships between the factors. Regardless of

choice, decisions should be supported by strong theoretical and mathematical justification, providing credibility to the final outcome" (Beavers et al., 2019). This technique can produce quality results when the sample size is less than 50 depending on the levels of loadings, number of factors, and number of variables (de Winter et al., 2009). The number of variables may not be an appropriate measure for determining the minimum sample size (Mundfrom et al., 2005). The participant-to-variable ratio may be more appropriate than the number of variables to determine sample size (Mundfrom et al., 2005). Mundfrom and colleagues' minimum necessary sample sizes table served as guidance for this study.

Beavers and colleagues provide several guidelines that we summarized in Table 11 (Beavers et al., 2019). This table is tailored to our study that explored EFA for validity analysis. We extracted a single factor using principal factor analysis to evaluate common variance and sums of squared loadings. We conducted four exploratory factor analyses using the relevance measure. Each construct varied in the number of variables: reliability (n=4), validity (n=4), internal validity (n=8), and external validity (n=4). The approximate participant-to-variable ratios varied by EFA: 5:1 (reliability), 5:1 (validity), 3:1 (internal validity), 5:1 (external validity). SPSS software version 28.0 was used for all statistical analyses.

Table 11: Adoption of Guidelines for Exploratory Factor Analysis from Beavers and Colleagues

Factorability

- *Correlation Matrix* – Review the correlation matrix. Correlations that exceed .30 provide sufficient evidence of commonality.
- *Determinant of the Matrix* – Generate determinant of the matrix. The determinant ranges from 0 to 1. In cases where the determinant is zero (<0.00001), there are no possible linear combinations or factors, indicating high multicollinearity among two or more variables. Non-zero determinants indicate that linear combinations exist. The smaller the determinant, the few linear combinations there are.
- *Bartlett's Test of Sphericity* – The Bartlett's Test of Sphericity is used to determine if the determinant is statistically different from zero.
- *Kaiser-Meyer-Olkin Test of Sampling Adequacy (KMO)* – The KMO provides the degree of common variance. We relied on KMO interpretation guidelines: marvelous (≥0.90), meritorious (0.80 – 0.89), middling (0.70 – 0.79), mediocre (0.60 – 0.69), miserable (0.50 – 0.59), and no factorability (≤0.50) (Kaiser, 1974).

Initial Extraction

- *Common Factor Analysis* – This study hypothesized an underlying construct for each EFA. Common factor analysis is more appropriate than principal component analysis (PCA) for this study because we assume that there is unique variance. Common factor analysis "removes specific variance and error variance from the calculations" (Beavers et al., 2019).
- *Method of Initial Extraction* – We used Principal Axis Factoring (PAF) because it does not make distributional assumptions such as normality and aligns with our research premise that there is an underlying construct. We have four constructs – reliability, validity, internal validity, and external validity.
- *First Factor* –The first factor accounts for the most variance.

Single-Factor Exploratory Analysis

- *Communalities* – The communality (h^2) or common variance is the proportion of variance for a given variable explained by the factor. It ranges from 0 (explains less of variance of item) – 1 (explains more of variance of item). Communality patterns can be defined as high (all values between .6 and .8), wide (all values between .2 and .8), and low (all values between .2 and .4)(Mundfrom et al., 2005).
- *Scree Plot* – This table provides the eigenvalues, which are not applicable to analysis.
- *Total Variance Explained* – This table provides the initial eigenvalue (assumes no unique variance), proportion of variance, and cumulative variance. Eigenvalues do not apply to our study because we did not conduct principal components analysis (PCA). We focus on the extraction values in this table on sums of squared loadings.
- *Item Correlation to Factor* – We compared the item correlations to the factor.
- *Factor Matrix* – This table provides the sums of squared loadings. We focus on this table because we sought sum of squared loadings (principal axis factoring). We follow the Costello and Osborne's rule of thumb, "If an item has a communality of less than .40, it may either a) not be related to the other items, or b) suggest an additional factor that should be explored" (Costello & Osborne, 2005).
- *Limitations* – This analysis is exploratory and may be apt to biased estimates. Data are interpreted considering strengths and limitations. The sample size from complete case analysis may not be adequate in all cases.

J. Benefits and Risks. There were direct benefits from participation in this research. New critical analysis skills for evaluating how race conceptualization, collection, and analysis can affect study quality were a potential benefit of participation. The Critical Race Framework itself was a benefit to improve personal research practices and analysis. The information that participants provided may help research better understand informed approaches for assessing the quality of research. The information that participants provided shed light on a section of research that needs further attention. Once participants have completed the training and survey, they will be entered into a drawing to receive monetary compensation. Five names were randomly drawn from among all study participants at the conclusion to receive one of five $100 gift cards.

Participants needed to provide their name and address to receive compensation if their name was drawn.

Any potential loss to confidentiality was minimized by keeping all data in a password protected computer application. All data were be kept in a secure location – behind password protected account using two-factor authentication. The University of Maryland storage site (UMD BOX) is highly secure and HIPAA compliant. Any report or article about this research project, protected participants' identity to the furthest extent possible. Participants' information could have been shared with representatives of the University of Maryland or governmental authorities if we observed that participants are in a life-threatening danger, or if we are required to do so by legal statutes. None of these procedures were necessary in this study. Only the principal investigator had access to the study data. Data from ineligible responses were destroyed before analysis in January 2024 and not used for research purposes.

6. Phase III

A. Method. Three researchers, including the PI, applied the Critical Race Framework 2.0 to 20 articles. These articles are derived from a) a sample of articles (n=10) from the bibliographic analysis of Arul and Mesfin (2017) and b) the ten articles from the search of highly cited behavioral health studies discussed above in Chapter 1 (Table 12) (Arias et al., 2003; Arul & Mesfin, 2017; Ashing-Giwa et al., 2004; Banks et al., 2006; Bell et al., 2018; Cuevas et al., 2020; De Hert et al., 2011; Devaraj et al., 2020; Haq & Penning, 2020; Luo et al., 2022; McClendon et al., 2021; Meissner et al., 2006; Mensah et al., 2005; Murray Horwitz et al., 2018; Parcha et al., 2020; Shavers & Brown, 2002; Skinner et al., 2003; Trivedi et al., 2005; Tuthill et al., 2020; Van Doorslaer et al., 2002). Ten articles from among 100 articles were selected using a random number generator using Google (*Google Random Number Generator - Google Search*,

n.d.). The article used in Phase I was excluded. The order was based on their appearance in the article. The random number generator produced the following results based on appearance in the article: 72, 16, 88, 69, 91, 55, 38, 37, 14, and 74.

Table 12: Article for Phase III – Application of Critical Race Framework	
Health Disparities Articles (n=10) *Assigned to Dr. Dyer (Rater 1)*	**Behavioral Health Articles (n=10)** *Assigned to Dr. Woodward (Rater 2)*
• Arias E, MacDorman MF, Strobino DM, et al. Annual summary of vital statistics—2002. *Pediatrics* 2003;112(6):1215–1230. (Appeared in list as #72) • Shavers VL, Brown ML. Racial and ethnic disparities in the receipt of cancer treatment. *Journal of the National Cancer Institute* 2002;94(5):334–357. (Appeared in list as #16) • Ashing-Giwa KT, Padilla G, Tejero J, et al. Understanding the breast cancer experience of women: a qualitative study of African American, Asian American, Latina and Caucasian cancer survivors. *Psycho-Oncology* 2004;13(6):408–428. (Appeared in list as #88) • Trivedi AN, Zaslavsky AM, Schneider EC, et al. Trends in the quality of care and racial disparities in Medicare managed care. *New England Journal of Medicine* 2005;353(7):692–700 (Appeared in list as #69) • Van Doorslaer E, Koolman X, Jones AM. Explaining income-related inequalities in doctor utilization in Europe. *Health Economics* 2004;13(7):629–647. (Appeared in list as #91) • Skinner J, Weinstein JN, Sporer SM, et al. Racial, ethnic, and geographic disparities in rates of knee arthroplasty among Medicare patients. *New England Journal of Medicine* 2003;349(14):1350–1359 (Appeared in list as #55) • Meissner HI, Breen N, Klabunde CN, et al. Patterns of colorectal cancer screening uptake among men and women in the United States. *Cancer Epidemiology Biomarkers & Prevention* 2006;15(2): 389–394. (Appeared in list as #38) • Banks J, Marmot M, Oldfield Z, et al. Disease and disadvantage in the United States and in England. *JAMA—Journal of the American Medical Association* 2006;295(17):2037–2045. (Appeared in list as #37) • Mensah GA, Mokdad AH, Ford ES, et al. State of disparities in cardiovascular health in the United States. *Circulation* 2005;111(10):1233–1241. (Appeared in list as #14) • De Hert M, Correll CU, Bobes J, et al. Physical illness in patients with severe mental disorders. I. Prevalence, impact of medications and disparities in	• Bell CN, Thorpe RJ, LaVeist TA. The role of social context in racial disparities in self-rated health. *Journal of Urban Health*. 2018;95(1):13-20. • Cuevas AG, Chen R, Slopen N, et al. Assessing the role of health behaviors, socioeconomic status, and cumulative stress for racial/ethnic disparities in obesity. *Obesity*. 2020;28(1):161-170. • Devaraj S, Stewart A, Baumann S, et al. Changes over time in disparities in health behaviors and outcomes by race in Allegheny County: 2009 to 2015. *Journal of Racial and Ethnic Health Disparities*. 2020;7(5):838-843. • Haq KS, Penning MJ. Social determinants of racial disparities in cognitive functioning in later life in Canada. *Journal of aging and health*. 2020;32(7-8):817-829. • Luo J, Hendryx M, Wang F. Mortality disparities between Black and White Americans mediated by income and health behaviors. *SSM-Population Health*. 2022;17:101019. • McClendon J, Chang K, Boudreaux MJ, Oltmanns TF, Bogdan R. Black-White racial health disparities in inflammation and physical health: Cumulative stress, social isolation, and health behaviors. *Psychoneuroendocrinology*. 2021;131:105251. • Murray Horwitz ME, Pace LE, Ross-Degnan D. Trends and disparities in sexual and reproductive health behaviors and service use among young adult women (aged 18–25 years) in the United States, 2002–2015. *American Journal of Public Health*. 2018;108(S4):S336-S343. • Parcha V, Malla G, Suri SS, et al. Geographic variation in racial disparities in health and coronavirus Disease-2019 (COVID-19) mortality. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*. 2020;4(6):703-716.

health care. *World Psychiatry* 2011;10(1):52–77. (Appeared in list as #74)	• Tuthill Z, Denney JT, Gorman B. Racial disparities in health and health behaviors among gay, lesbian, bisexual and heterosexual men and women in the BRFSS-SOP. *Ethn Health*. 2020;25(2):177-188. • Whitaker KM, Jacobs Jr DR, Kershaw KN, et al. Racial disparities in cardiovascular health behaviors: the coronary artery risk development in young adults study. *American journal of preventive medicine.* 2018;55(1):63-71.
Note: These ten articles in the first column were derived from a sample of articles from the bibliographic analysis of Arul and Mesfin (2017) and b) the ten articles from a Google Scholar search of behavioral health studies between 2018-2022, described in Chapter 2 (Arul & Mesfin, 2017).	

Table 13 briefly discusses Steps 13-14 for the CR Framework conceptual model. As discussed, three raters scored 20 articles – 10 for each co-research and 20 for the PI. No additional data were collected. We discuss our reliability analysis further in the analysis section (I. Analysis).

B. Design. The design of Phase III is a series of article evaluations.

C. Study Sample. The samples for Phase III were research articles from the health disparities and behavioral health literature. The sampling strategy is described in the methods section above (Table 10).

D. Sample Size. Phase III did not involve human subjects research. Twenty (20) articles were analyzed. Ten was the maximum of articles that each co-researcher could realistically evaluate during the study period. A sample of ten studies has been used in prior critical appraisal studies (Kmet et al., 2004).

E. Recruitment. The PI recruited two co-researchers for this phase of the study. Two coders are common and acceptable for assessing interrater reliability and methodological quality of a critical appraisal tool (Hallgren, 2012; Pace et al., 2012; Pluye et al., 2009). The study design and coders are justified because determining whether measures are theoretically linked is usually based on qualitative evaluation among experts (Weiner et al., 2017). Dr. Typhanye Dyer is an Associate Professor of Epidemiology and Biostatistics at the University of Maryland (UMD) – College Park. She is an epidemiologist with a focus on health disparities – STI and HIV

Table 13: Conceptual Framework for CR Framework Development (Phase III – 15-14)

- *Step 13. Article Evaluation* – Following refinement in Phase II, three researchers applied the CR Framework. Each co-researcher evaluated 10 articles (Table 12). The PI (Rater 3) evaluated all 20 articles.
- *Step 14. Tests of Reliability* – We assessed interrater agreement and reliability testing.

scholarship. Dr. Nathaniel Woodward is a postdoctoral research fellow at the University of North Carolina at Chapel Hill Lineberger Comprehensive Cancer Center.

F. Procedures. The principal investigator (Rater 3) downloaded each of the 20 articles. In early January 2024, he provided co-researchers Dr. Typhanye Dyer (Rater 1) and Dr. Nathaniel Woodard (Rater 2) with a file containing each of their set of ten articles, the Critical Race Framework, and a response sheet. They were also able to provide their responses online via the study website (www.criticalraceframework.com). Phase III occurred throughout the month of January. Dr. Dyer rated the health disparities articles. Dr. Woodward rated the behavioral health articles. No additional data were collected in Phase III.

G. Data Collection. Each reviewer provided responses to each question in the CR Framework for 10 articles.

H. Instruments. The PI used the Critical Race Framework 2.0.

I. Analysis.

1. Interrater Agreement. We calculated interrater agreement as percentage agreement using two methods: absolute agreement and high/moderate versus low/no discussion by item and rater pair.

2. Interrater Reliability. We used the weighted kappa (quadratic weights) for interrater reliability analysis using absolute agreement. We did not conduct interrater reliability analysis for high/moderate versus high/low because our data were not numerous and varied for testing (e.g., constants present for one or both raters).

J. Benefits and Risks. Phase III involved two co-researchers who performed article critiques. There were direct benefits for the co-researchers. The PI will invite them to participate in manuscript development and co-authorship following dissertation defense. The PI also offered voluntary research hours for data cleaning and analysis. The risks to the co-researchers are minimal. It is possible that 10 article critique can induce survey fatigue or stress. This study mitigated this by providing nearly a full month for task completion.

Summary of Chapter 3

Chapter 3 discussed the three phases of our study to develop and test the Critical Race Framework critical appraisal tool, including our study design, methodology, instruments, sampling strategies, and analytical plans. In Phase I, we pilot-tested the tool and gathered feedback to improve the training, instrument, and study design and to assess our *measures of fit*.

In Phase II, we conducted validity and *measures of fit* analyses from a national sample of public health experts using the third version of the tool.

Chapter 4: Results

1. Introduction to Chapter 4

The Critical Race Framework study recruited 30 highly skilled public health experts (Phase 1, n=6; Phase 2, n=22; Phase 3, n=2) who informed study design, instrumentation, and training. Their quantitative and qualitative data were used to answer five research questions. This chapter provides our data results by phase and research question.

2. Results of Phase I

Approximately 280 faculty and PhD students were contacted to participate in Phase I. Sixteen individuals enrolled in the study, resulting in a 5% response rate. The enrollment cap of fifteen (15) was reached on June 18, 2023. Eight participants (50%) completed Part I – Critical Race Framework training. Six participants (38%) completed Part II – article critique and survey. The attrition rate across the entire study was high at 63% (10 out of 16). All study enrollees answered the immediately preceding question on whether they had used a variable for race in their research. If they answered "yes" to this question (n=14), then they were asked to discuss personal assumptions about race. All 14 respondents answered this question. Seven users did not proceed with the study after this question, accounting for 70% of attrition in the study. Qualtrics user data analytics indicated that the page containing the introductory video on the Critical Race Framework training was not displayed to any of these seven users, meaning that they answered the last question but did not click the arrow button to proceed. Nine participants downloaded and read the Critical Race Framework. Among these, one participant did not complete Part 1. Two participants who completed Part 1 did not return for Part 2.

The missingness rate was 59% (843 out of 1440). We concluded that the data were MNAR – that there were unobservable factors that directly affected missingness. Study

participants were presented with the purpose of the study ("The purpose of this research project is to assess the quality of a critical analysis tool for evaluating health research studies that use a racial variable") during enrollment. The Critical Race Framework itself was an unlikely explanation because 70% did not see the training or tool.

Attrition was higher among those with a master's degree (80%) compared to a PhD (55%), but PhD-holders made up more of the dropout population (60%). PhD-holders comprised 83% of the analytic sample (Table 14). Doctoral students (n=5) had the highest attribution rate (80% vs. 50%) compared to teaching faculty (n=8). Non-teaching faculty (n=1) had no attrition. Faculty comprised five of the six analytic samples. The percentage of individuals with high familiarity ("very familiar") remained relatively stable throughout the study.

Table 14: Demographics and Perceptions By Part in Phase I

	Enrollees n (%)	**Part 1 Completion n (%)**	**Part 2 Completion n (%)**
Total (n, %)	16 (100)	8 (50)	6 (75)
Participation Rate			
Between initial and second solicitations[1]	1 (6)	1 (13)	1 (17)
After Second Solicitation[1]	15 (94)	7 (88)	5 (83)
Next Component Completion	8 (50)[1]	6 (75)	
Highest Educational Attainment			
PhD or equivalent	11 (69)	5 (63)	5 (83)
Master's or equivalent	5 (31)	3 (38)	1 (17)
Current role			
Teaching faculty	8 (50)	4 (50)	4 (67)
Doctoral students	5 (31)	3 (38)	1 (17)
Non-teaching faculty	1 (6)	1 (13)	1 (17)
Familiary[2]	**n (%)**	**n (%)**	**n (%)**
Reliability			
Very familiar	11 (69)	6 (75)	4 (67)
Somewhat/fairly	5 (31)	2 (25)	2 (33)
Validity			
Very familiar	11 (69)	6 (75)	4 (67)
Somewhat/fairly	5 (31)	2 (25)	2 (33)
Internal validity			
Very familiar	10 (63)	6 (75)	4 (67)
Somewhat/fairly	6 (38)	2 (25)	2 (33)
External validity			
Very familiar	10 (63)	6 (75)	4 (67)
Somewhat/fairly	6 (38)	2 (25)	2 (33)
No. of years public health research (n=16)			

None	1 (6)	0 (0)	—
<1	0	0	
1-3	0	0	
4-6	6 (38)	3 (38)	2 (33)
7-15	4 (25)	1 (13)	—
16-25	0	0	
26-40	4 (25)	3 (38)	3 (50)
16-25	1 (6)	1 (13)	1 (17)
>25	0	0	
Five most common social Identities (n=16)[3]			
Woman	4 (25)	3 (38)	2 (33)
Male	3 (19)	3 (38)	3 (50)
Mother	3 (19)	0 (0)	—
Professor	3 (19)	0 (0)	—
Researcher	3 (19)	2 (25)	2 (33)

[1]Initial study solicitations were sent on May 2, 2023, and second solicitations were sent on May 13, 2023.
[2]Scale 1-5: Not at all familiar (excluded from study), somewhat familiar, fairly familiar, very familiar.
[3]Categories are based on most common from pool of study enrollees. This was an open-ended question (comments box), "What are your most important social identities? Social identity is a person's sense of who they are based on their group membership(s)."

RQ1. To what extent is the CR Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs? (Phase I).

A. Measures of Fit

A.1) Acceptability

The CR Framework exceeded the threshold for acceptability (≥60%) for each scale item (Table 15). Sixty-seven percent (n=4) of respondents agreed or completely agreed on four of the five scale items. The framework was "appealing" for 5 of the six respondents.

Table 15: Phase I Acceptability Results (n=6)

Scale Items	Completely disagree/disagree	Neither agree nor disagree	Completely Agree/Agree
Meets approval		2 (33)	4 (67)
Appealing		1 (17)	5 (83)
"I welcome"		2 (33)	4 (67)
No objection	1 (17)	1 (17)	4 (67)
"I like"		2 (33)	4 (67)

Note: Scale questions appeared as follows: "Critical Race Framework meeting my approval," "Critical Race Framework is appealing to me," "I welcome the Critical Race Framework," "I have no objection to the Critical Race Framework," and "I like the Critical Race Framework." The Likert scale included options for completely disagree, disagree, neither agree nor disagree, agree, and completely agree.

A.2) Satisfaction

The CR Framework was satisfactory ("satisfied" or "strongly satisfied") for more than 60% of respondents (Table 16). Satisfaction was assessed in Part 1. Three participants were neutral in satisfaction. No participant indicated dissatisfaction.

Table 16: Phase I Satisfaction Results (n=8)			
	Dissatisfied/Strongly Dissatisfied	**Neutral**	**Satisfied/Strongly Satisfied**
Satisfaction with Phase I training	0 (0)	3 (38)	5 (63)
Note: Satisfaction was assessed during Part 1 using a single-item Likert scale, "Please rate your satisfaction with the Phase I training". The options were strongly disagree, disagree, neutral, agree, strongly agree. Collapsed data were reported (strongly disagree/disagree, neutral, agree/strongly agree).			

A.3) Feasibility

We directly and indirectly assessed feasibility – scale and user data analytics on duration. Feasibility of the CR Framework training and study design was favorable based on the scale. It did not achieve feasibility due to the length of time spent on study activities. Overall, at least 60% of respondents indicated that the CR Framework was feasible based on item-level analysis (Table 17). Three items garnered complete agreement – "possible," "doable," and "workable." There was an even split in whether the framework was easy to use – disagreement (50%) and agreement (50%).

Table 17: Phase I Feasibility Results (n=6)			
Scale Items	**Completely disagree/disagree**	**Neither agree nor disagree**	**Completely Agree/Agree**
Implementable	1 (17)	1 (17)	4 (67)
Possible			6 (100)
Doable			6 (100)
Easy to use	3 (50)		3 (50)
Workable			6 (100)
Note: Scale questions appeared as follows: "The Critical Race Framework seems implementable," "The Critical Race Framework seems possible," "The Critical Race Framework seems doable," "The Critical Race Framework seems easy to use," and "The Critical Race Framework seems workable." The Likert scale included the following options: completely disagree, disagree, neither agree nor disagree, agree, and completely agree.			

Seven (88%) of Part 1 participants spent 21-40 minutes completing the training (Table 18). Another 21-40 minutes were spent each on reading and understanding the Critical Race Framework for seven participants and completing the framework for the assigned article. System-generated times indicated that most participants spent 30-60 minutes on Part 1 (63%) and less than 30 minutes for Part II (67%). Half of respondents (n=3) spent 21-40 minutes reading the article. Roughly one-third (n=2) spent between 41-60 minutes. Taken together, most participants fell in a range of 1.5-2.5 hours for all activities in Phase I. It exceeded the anticipated completion time (45 minutes).

Table 18: Phase I Duration Results

	Part 1 Survey (n=8)	**Part 2 Survey (n=6)**	**Very Familiar[3] (n=6)**	**Role**
Training[1]	**n (%)**	**n (%)**	**n**	
11-20 mins	0 (0)	—		
21-40 mins	7 (88)	—	3	Teaching faculty (n=3) Doctoral student (n=1) Non-teaching faculty (n=1)
41-60 mins	1 (13)	—	1	Teaching faculty (n=1)
Read and Understand CR Framework[2]	**n (%)**	**n (%)**	**n**	
>11 mins	—	1 (17)	1	Teaching faculty (n=1)
11-20 mins	—	0 (0)	—	—
21-40 mins	—	4 (67)	3	Teaching faculty (n=2) Doctoral student (n=1)
41-60 mins	—	1 (17)	1	Non-teaching faculty (n=1)
Reading Article[3]	**n (%)**	**n (%)**	**n**	Role (total)
11-20 mins	—	1 (16)	1	Teaching faculty (n=1)
21-40 mins	—	3 (50)	2	Teaching faculty (n=2) Doctoral student (n=1)
41-60 mins	—	2 (33)	1	Teaching faculty (n=1) Non-teaching faculty (n=1)
CR Framework Completion[4]	**n (%)**	**n (%)**	**n**	
11-20 mins	—	1 (17)	1	Teaching faculty (n=1)
21-40 mins	—	5 (83)	3	Teaching faculty (n=3) Doctoral student (n=1) Non-teaching faculty (n=1)
41-60 mins	—	0 (0)		
System-Generated Completion Time[4]	**n (%)**	**n (%)**	**n**	
<30 mins	1 (13)	4 (67)	4	Teaching faculty (n=4)
30 – 60 mins	5 (63)	1 (17)	0	Doctoral student (n=1)
>60 mins	2 (25)	1 (17)	0	Non-teaching faculty (n=1)

Average number of days (>60 mins only)	5.8	0 (0)	0	Non-teaching faculty (n=1)
Time between end of Part 1 and start of Part 2 (days)	**n (%)**	**n (%)**	**n**	
< 1 day	—	2 (50)	2	Teaching faculty (n=2)
1 -2 day(s)	—	1 (17)	1	Teaching faculty (n=1)
5 days		1 (17)	1	Teaching faculty (n=1)
8 days	—	1 (17)	0	Doctoral student (n=1)
21 days	—	1 (17)	0	Non-teaching faculty (n=1)

1 This question appeared in Part 1, "Approximately how long did it take to complete the Critical Race Framework training from start to finish, including reading the CR Framework and watching the videos?"
2 Approximately, how long did it take for you to read and fully understand the Critical Framework?"
3 "Approximately, how long did it take for you to read and fully understand the article, Eight Americas: Investigating Mortality Disparities across Races, Counties, and Race-Counties in the United States by Christopher J. L Murray and colleagues? Please do not include the time spend completing the CR Framework for this article."
4 Approximately, how long did it take for you to complete the Critical Framework for the article?
5 Complete responses only (n=8). Qualtrics defines duration a) for complete responses as the time from start to finish, including breaks and b) for incomplete responses as the time from start to the completion data based on survey settings.
3 Number of respondents with self-reported of "very familiar" to four core concepts.

Appropriateness

A.3) Appropriateness was directly and indirectly assessed. Overall, more than 60% of participants agreed on the appropriateness of the CR Framework (Table 19). 50% (n=3) agreed that the framework was "well-aligned". An equal percentage was neutral ("neither agree nor disagree"). A single participant disagreed that the framework was suitable.

Table 19: Phase I Appropriateness Results (n=6)

Scale Item	Completely disagree/disagree	Neither agree nor disagree	Completely Agree/Agree
Fitting	0	2 (33)	4 (67)
Suitable	1 (17)	1 (17)	4 (67)
Applicable	0	1 (17)	5 (83)
Good match	0	2 (33)	4 (67)
Well-aligned	0	3 (50)	3 (50)

We also assessed appropriateness based on responses on relevance. This question asked respondents to indicate the relevance of each question for the related construct (Table 20). All questions in the framework were rated fairly or highly relevant to their related construction by more than 60% of participants. The single exception is "Discussed whether a "true value" exists

for race". Fifty percent felt that this question was somewhat relevant to internal validity. An equal percentage rated it as fairly or highly relevant. The percentage of "highly relevant" responses for each construct is as follows: reliability (50%, 9 out of 18 possible response), validity (39%, 14 out of 36 possible response), internal validity (45%, 49 out of 108 possible responses, and external validity (50%, 9 out of 18 responses). The percentage of "somewhat relevant" or "highly relevant" responses for each construct is as follows: reliability (83%, 15 out of 18 possible response), validity (78%, 28 out of 36 possible response), internal validity (80%, 86 out of 108 possible responses, and external validity (83%, 15 out of 18 responses).

Table 20: Phase I Relevance Results

	Not at All Relevant	Somewhat Relevant	Fairly Relevant	Highly Relevant
Relevance to Reliability				
Discussed reliability of racial measurement tool?	0	1 (17)	2 (33)	3 (50)
Cited prior reliability evidence?	0	1 (17)	2 (33)	3 (50)
Supported reliability of racial measurement tool with multiple measurement?	0	1 (17)	2 (33)	3 (50)
Relevance to Validity				
Discussed what construct or meaning race intended to measure?	0	1 (17)	2 (33)	3 (50)
Discussed relevance of racial admixture to construct or meaning of race used in study?	0	1 (17)	2 (33)	3 (50)
Discussed whether a "true value" exists for race?	0	2 (33)	3 (50)	1 (17)
Cited validity evidence for racial measurement tool?	0	1 (17)	2 (33)	3 (50)
Discussed characteristics intended to distinguish racial groups?	0	2 (33)	2 (33)	2 (33)
Assessed whether racial measurement tool correctly and consistently distinguished races?	0	1 (17)	3 (50)	2 (33)
Relevance to Internal Validity				
Assessed threats to internal validity due to race variable?	0	1 (17)	2 (33)	3 (50)
Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	0	1 (17)	3 (50)	2 (33)
Discussed limitations of race to establish quality of study design?	0	1 (17)	2 (33)	3 (50)
Cited theories and evidence to inform how race is treated methodologically and support by theory of change?	0	2 (33)	3 (50)	1 (17)
Discussed potential participant sources of measurement error related to racial measurement tool?	0	0	3 (50)	3 (50)
Discussed potential sources of measurement error related to racial measurement tool?	0	0	3 (50)	3 (50)

Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool?	0	1 (17)	2 (33)	3 (50)
Discussed limitations of assessing measurement error due to an unknown "true value" for race?	0	3 (50)	1 (17)	2 (33)
Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other,")	0	1 (17)	2 (33)	3 (50)
Reported rate of racial switching or changes in participant racial classification?	0	2 (33)	1 (17)	3 (50)
Justified combining survey options or combinations of race variable in analysis that differ from instrument?	0	1 (17)	1 (17)	4 (67)
Discussed racial subgroup analysis?	0	2 (33)	2 (33)	2 (33)
Discussed testing of statistical assumptions and implications for race?	0	1 (17)	4 (67)	1 (17)
Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	0	1 (17)	2 (33)	3 (50)
Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	0	2 (33)	1 (17)	3 (50)
Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities?	0	1 (17)	2 (33)	3 (50)
Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	0	1 (17)	1 (17)	4 (67)
Discussed implications for internal validity due to treatment of race variable analytically?	0	1 (17)	2 (33)	3 (50)
Relevance to External Validity				
Discussed threats to external validity considering race-related threats to internal validity?	0	1 (17)	2 (33)	3 (50)
Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	0	1 (17)	3 (50)	2 (33)
Discussed other threats to external validity for generalizability claims related to race?	0	1 (17)	1 (17)	4 (67)

B. Potential Threats to Reliability

We analyzed the data set for potential response bias due to non-differentiation or straight-lining. Table 21 displays the results using the simple non-differentiation method. Possible values range from 0 – 1.0. The most common values (0, 0.33, and .67) tied for three occurrences. A single rater had "yes" for all items in the Critical Race Framework during the article critique phase of the study. This pattern appeared for no other user.

Table 21 - Results of Simple Non-Differentiation Method by *Measures of Fit* and Article Evaluation Responses Using the Critical Race Framework		
Measure	**N**	**Values (Proportion)**
Overall *Measures of Fit* (1-3)	0	0
1. Acceptability	3	0.5
2. Feasibility	2	0.33
3. Appropriateness	3	0.5
Relevance (Overall)	0	0
Reliability	6	1.0
Validity	4	0.67
Internal Validity	0	0
External Validity	4	0.67
CR Framework Evaluation (overall)	1	.17
Reliability	4	.67
Validity	1	.17
Internal Validity	2	.33
External Validity	2	.33
Note: The simple nondifferentiation measure at the scale-level by calculating the percentage of study respondents who used an identical response.		

Questions on data question clarity, participant understanding, and redundancy were not entirely reliable due to a technical error that forced a response for some users. The study had intended this question to be optional, where appropriate. This forced response feature was turned off once the PI was alerted to this error. It appeared to have affected half of Part 2 participants. A single affected user provided their accurate responses in the final open comments box ("14. redundant 17. reword 21 reword"). However, we were unable to determine how to modify the count of the original data since we do not know their original response. We retained this user's data for refinement decisions.

"Should be reworded for clarity" appeared in the first column in the matrix table on the survey. While it is likely that the survey forced inaccurate responses (e.g., none of the options were applicable) for some, an alternative interpretation is that the survey forced respondents to choose one of three areas of improvement that could serve as the basis for refinement of the CR Framework. Table 22 displays the results as provided. Responses with 4 or more indications were most common for internal validity (9 out of 11)(Table 22). Internal validity had 27 items in

the Critical Race Framework. Three participants had difficulty understanding 16 questions across all core constructs.

Table 22: Indications of Improvement (n=6)				
No.	CR Framework	Should be reworded for clarity (n)[1]	Difficult to understand (n)[1]	Redundant (n)[1]
Reliability				
1.	Discussed reliability of racial measurement tool?	2	1	
2.	Cited prior reliability evidence?	3		
3.	Supported reliability of racial measurement tool with multiple measurement?	2	1	
Validity				
4.	Discussed what construct or meaning race intended to measure?	3		
5.	Discussed relevance of racial admixture to construct or meaning of race used in study?	3		
6.	Discussed whether a "true value" exists for race?	2	2	
7.	Cited validity evidence for racial measurement tool?	3		
8.	Discussed characteristics intended to distinguish racial groups?	1	1	1
9.	Assessed whether racial measurement tool correctly and consistently distinguished races?	2	1	
Internal Validity				
10.	Assessed threats to internal validity due to race variable?	4		
11.	Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	4		
12.	Discussed limitations of race to establish quality of study design?	4		
13.	Cited theories and evidence to inform how race is treated methodologically and support by theory of change?	3	1	
14.	Discussed potential participant sources of measurement error related to racial measurement tool?	4		1
15.	Discussed potential sources of measurement error related to racial measurement tool?	4		
16.	Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool?	3		1
17.	Discussed limitations of assessing measurement error due to an unknown "true value" for race?	5		
18.	Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other,")	4		
19.	Reported rate of racial switching or changes in participant racial classification?	3	1	
20.	Justified combining survey options or combinations of race variable in analysis that differ from instrument?	3	1	

21.	Discussed racial subgroup analysis?	3	2	
22.	Discussed testing of statistical assumptions and implications for race?	2	1	1
23.	Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	4		
24.	Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	3	1	
25.	Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities?	4		
26.	Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	3	1	
27.	Discussed implications for internal validity due to treatment of race variable analytically?	3		1
External Validity				
28.	Discussed threats to external validity considering race-related threats to internal validity?	4		
29.	Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	4		
30.	Discussed other threats to external validity for generalizability claims related to race?	2	2	

RQ2. What improvements to the Critical Race Framework 1.0 study design, instrumentation, and training are needed? (Phase I).

In conjunction with the previous study findings (RQ1), we relied on Critical Race Framework article critiques and qualitative data results to identify areas of improvement in study design, instrumentation, and training. Attrition was 63% (10 of out 16 study enrollees). Table 23 provides summary data on article critique using the Quality of Evidence scale – a four-item quality scale on each domain. Most study participants rated the article "very low" or "low" quality in every area (n=4). Table 24 provides critical evaluation ratings using the CR Framework. This table includes the PI's scores and adjusted agreement percentage. Table 25 provides the PI's rationale for his rating. Of the 13 instances of majority agreement (>50%), the PI agreed 69% of the time (9 out of 13). The PI rated the article before viewing any survey data.

Table 23: Quality of Evidence Results				
	Very low n=6	**Low n=6**	**Moderate n=6**	**High n=6**
Reliability Evidence	3 (50)	1 (17)	2 (33)	0 (0)
Validity Evidence	3 (50)	1 (17)	2 (33)	0 (0)
Internal validity evidence	3 (50)	1 (17)	2 (33)	0 (0)
External validity evidence[4]	3 (50)	1 (17)	2 (33)	0 (0)
Note: Four questions about quality of evidence in reliability, validity, internal validity, and external validity using four-item scale (very low, low, moderate, and high) - limited to racial variable analysis				

Table 24: Critical Race Framework 1.0 Ratings						
No.	**CR Framework**	**Yes** n (%)	**No** n (%)	**Unclear** n (%)	**N/A** n (%)	**With PI ("No")** n (%)
Reliability						
1.	Discussed reliability of racial measurement tool?	3 (50)	3 (50)	0 (0)	0 (0)	4 (57)
2.	Cited prior reliability evidence?	2 (33)	4 (67)	0 (0)	0 (0)	5 (71)
3.	Supported reliability of racial measurement tool with multiple measurement?	2 (33)	4 (67)	0 (0)	0 (0)	—
Validity						
4.	Discussed what construct or meaning race intended to measure?	2 (33)	3 (50)	1 (17)	0 (0)	4 (57)
5.	Discussed relevance of racial admixture to construct or meaning of race used in study?	3 (50)	2 (33)	1 (17)	0 (0)	3 (43)
6.	Discussed whether a "true value" exists for race?	1 (17)	4 (67)	1 (17)	0 (0)	5 (71)
7.	Cited validity evidence for racial measurement tool?	2 (33)	3 (50)	1 (17)	0 (0)	4 (57)
8.	Discussed characteristics intended to distinguish racial groups?	4 (67)	1 (17)	1 (17)	0 (0)	2 (29)
9.	Assessed whether racial measurement tool correctly and consistently distinguished races?	3 (50)	3 (50)	0 (0)	0 (0)	4 (57)
Internal Validity						
10.	Assessed threats to internal validity due to race variable?	3 (50)	3 (50)	0 (0)	0 (0)	4 (57)
11.	Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	4 (67)	1 (17)	1 (17)	0 (0)	—
12.	Discussed limitations of race to establish quality of study design?	5 (83)	1 (17)	0 (0)	0 (0)	2 (29)
13.	Cited theories and evidence to inform how race is treated methodologically and support by theory of change?	1 (17)	4 (67)	1 (17)	0 (0)	—
14.	Discussed potential participant sources of measurement error related to racial measurement tool?	4 (67)	1 (17)	1 (17)	0 (0)	—

15.	Discussed potential sources of measurement error related to racial measurement tool?	5 (83)	1 (17)	0 (0)	0 (0)	—
16.	Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool?	2 (33)	2 (33)	2 (33)	0 (0)	—
17.	Discussed limitations of assessing measurement error due to an unknown "true value" for race?	2 (33)	3 (50)	1 (17)	0 (0)	4 (57)
18.	Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other,")	2 (33)	3 (50)	1 (17)	0 (0)	4 (57)
19.	Reported rate of racial switching or changes in participant racial classification?	3 (50)	3 (50)	0 (0)	0 (0)	4 (57)
20.	Justified combining survey options or combinations of race variable in analysis that differ from instrument?	2 (33)	1 (17)	0 (0)	3 (50)	—
21.	Discussed racial subgroup analysis?	5 (83)	1 (17)	0 (0)	0 (0)	—
22.	Discussed testing of statistical assumptions and implications for race?	1 (17)	4 (67)	1 (17)	0 (0)	5 (71)
23.	Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	1 (17)	4 (67)	1 (17)	0 (0)	5 (71)
24.	Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	1 (17)	5 (83)	0 (0)	0 (0)	—
25.	Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities?	2 (33)	2 (33)	2 (33)	0 (0)	3 (43)
26.	Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	1 (17)	3 (50)	2 (33)	0 (0)	—
27.	Discussed implications for internal validity due to treatment of race variable analytically?	3 (50)	3 (50)	0 (0)	0 (0)	4 (57)
External Validity						
28.	Discussed threats to external validity considering race-related threats to internal validity?	3 (50)	3 (50)	0 (0)	0 (0)	—
29.	Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	2 (33)	3 (50)	1 (17)	0 (0)	—
30.	Discussed other threats to external validity for generalizability claims related to race?	3 (50)	3 (50)	0 (0)	0 (0)	—
	Average	2.6 (43)	2.7 (45)			

Table 25 displays the principal investigator's response and rationale for each rating in the Phase I article critique. His evaluation found 17 "yes", 10 "no", and 3 "not applicable" ratings.

Table 25 - Principal Investigator's Rationale for Critical Race Framework Ratings			
No.	**CR Framework**	**Response**	**Justification**
1.	Discussed reliability of racial measurement tool?	No	No quality information on reliability given associated with race data collection tool in 1) race categories in National Center for Health Statistics (NCHS) mortality files and 2) census.
2.	Cited prior reliability evidence?	No	They cited no sources to support reliability of the racial data collection tool.
3.	Supported reliability of racial measurement tool with multiple measurement?	Not applicable	The study did not conduct any primary data collection.
4.	Discussed what construct or meaning race intended to measure?	No	The authors never explain the construction of race.
5.	Discussed relevance of racial admixture to construct or meaning of race used in study?	No	No, the study in fact defaults to 1977 definition rather than more inclusive definition from 1997.
6.	Discussed whether a "true value" exists for race?	No	No, study strongly implies that there is a true value.
7.	Cited validity evidence for racial measurement tool?	No	There is no validity evidence.
8.	Discussed characteristics intended to distinguish racial groups?	No	No discussion
9.	Assessed whether racial measurement tool correctly and consistently distinguished races?	No	No discussion
10.	Assessed threats to internal validity due to race variable?	No	The authors discuss 1) back-migration among Asians impacting the results of America 1, 2)
11.	Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	Not applicable	There was no treatment in this study.
12.	Discussed limitations of race to establish quality of study design?	No	Study assumes primacy of race in study quality.
13.	Cited theories and evidence to inform how race is treated methodologically and support by theory of change?	Yes	Citations 25, 27, 29, and 30 in article support rating. Study relies on common practices for methodological treatment of race.
14.	Discussed potential participant sources of measurement error related to racial measurement **tool**?	Yes	Yes, acknowledged that physician or facility may be source of error. "For Native Americans, misreporting race on death certificates is most important where mixed races exist and Native Americans form a small proportion of the population (e.g., in California)."
15.	Discussed potential sources of measurement error related to racial measurement tool?	Yes	Discussed, "In these groups, it is more likely for race to be reported as Native American or Asian in the census than it is in the death certificate; hence, the differential underestimation of deaths and population results

			in bias in the form of a net underestimation of mortality"
16.	Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool?	Yes	A secondary source of bias for Asians may be back-migration of first-generation immigrants, who return to their home countries due to illness.
17.	Discussed limitations of assessing measurement error due to an unknown "true value" for race?	No	Implied assumption that there is a true value
18.	Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other,")	No	Major collapsing of data; results are in bridged data methods, but not reported in study
19.	Reported rate of racial switching or changes in participant racial classification?	No	No.
20.	Justified combining survey options or combinations of race variable in analysis that differ from instrument?	Yes	Yes, but this is a value judgment. Their justification was that NCHS combined in this way because 1990 and 2000 race surveys were different
21.	Discussed racial subgroup analysis?	Yes	They held on to race, but used different Americas
22.	Discussed testing of statistical assumptions and implications for race?	No	Conducted tests for life expectancy and probabilities of death, but provided no data on
23.	Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	No	There are aspects of the race variable that it looked at such as Asian, but this discussion was not adequate
24.	Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	Yes	"For Asians, this source of bias was addressed by adjusting population and mortality for differential underestimation using the National Mortality Follow-back Survey" I do not think it was an adequate adjustment.
25.	Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities?	No	They discussed racial gaps in life expectancy and morality. They do not consider whether study subjects are independent. Is the life expectant analysis if parametric or nonparametric.
26.	Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	NA	None of the analyses involved statistical testing (testing by the authors, not outside of the study)
27.	Discussed implications for internal validity due to treatment of race variable analytically?	No	No. The authors assumes that the methods for Census and bridged data are high quality.
28.	Discussed threats to external validity considering race-related threats to internal validity?	Yes	Authors assume the primacy of racial scientific reasoning, but discuss several limitations: "reported race in the census, used for population estimates, may be different from race in mortality statistics, where race may be reported by the family, the…" "it is more likely for race to be reported as Native American or Asian in the census than it is in the death certificate; hence,

			the differential underestimation of deaths and population results in bias in the form of a net underestimation of mortality"
29.	Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	Yes	Yes, to the extent that different races are in different "Americas"
30.	Discussed other threats to external validity for generalizability claims related to race?	Yes	Yes, "A secondary source of bias for Asians may be backmigration of first-generation immigrants, who return to their home countries due to illness." However, this question is too open-ended.

Table 26 provides qualitative data from Phase I. Comments were organized into one of three *a priori* themes: training, instrument, and study design. The table also includes the self-ratings in familiarity of which there were four questions. All raters had the same self-assessed familiarity (e.g., "fairly familiar" for each self-rating). Defined as the web-based video and transcript, *training* received the most comments (n=4, 50%). Feedback was both constructive ("The training was too long; Could not enlarge the video or move it to follow along with the transcript") and commending ("It was clear, concise, informative and advanced, at the same time"). The *instrument* or Critical Race Framework garnered two comments from separate individuals. One was highly detailed in constructive feedback on the scale, word selection, and length, "Be careful of double-barreled items. Consider a scale (e.g., 1-5, instead of dichotomous measures)? Be careful and intentional about discuss vs assess and communicate to rater what the difference is. In some instances, it seemed to me that authors may have discussed, though not assessed, a certain item…" The other comment expressed benefits of the framework, "This is a very detailed instrument for practitioners to use regularly." The study design received constructive feedback, "I was highly impressed by this work but then I got lost" and "The last page of the survey was flawed and required responses to all questions in order to submit the survey."

Table 26: Phase I Qualitative Data				
No.	**Comments**	**Self-Reported Familiarity Rating[1]**	**Role**	**Primary Qualitative Theme**
Part 1				
1.	"I think this training should be replicated. It was clear, concise, informative and advanced, at the same time. Timing of slides and the narration were done very well."	Very familiar (n=4)	Teaching faculty	Training
2.	"This is a very detailed instrument for practitioners to use regularly. The principles make sense and will remind readers of the many limitations (and) weaknesses inherent in using race categories in any study, but the full blown use of the tool will likely be limited to individuals seeking to study the use of race in research itself."	Very familiar (n=4)	Teaching faculty	Instrument
3.	"The training was too long; Could not enlarge the video or move it to follow along with the transcript. The transcript was not verbatim."	Very familiar (n=4)	Teaching faculty	Training
4	"Consider a more interactive training."	Fairly familiar (n=4)	Doctoral student	Training
5.	"The videos were helpful and clearly explained each section of the CR Framework. It was helpful to have transcripts available. This approach takes into consideration that people have different learning styles."	Fairly familiar (n=4)	Non-teaching faculty	Training
Part 2				
6.	"I was highly impressed by this work but then I got lost. The modules were clear but the application of the CR framework to what is basically a descriptive epidemiology paper was confusing. I would hav(e) much rather had an empirical paper for which measures, and discussion of analyses and threats, bias, etc were actually discussed. Maybe I misinterpreted what I was supposed to do. But I then found the survey afterwards to be really confusing, as I was not sure if the questions were in relation to the CR framework AND the paper or just the VR framework or both? That's when things got really unclear."	Very familiar (n=4)	Teaching faculty	Study design
7.	"The last page of the survey was flawed and required responses to all questions in order to submit the survey. I had to submit answers for all of them, when it was supposed to only be those I wanted to identify as needing some revision. I only wanted to select the following questions as benefiting from revision. All of the other questions were good. 14. redundant 17. reword 21 reword"	Very familiar (n=4)	Teaching faculty	Study design
8.	"Be careful of double-barreled items Consider a scale (e.g., 1-5, instead of dichotomous measures)? Be careful and intentional about discuss vs assess and communicate to rater what the difference is. In some instances it seemed to me that authors may have discussed, though not assessed, a certain item. How the rater reads and interprets the difference between discuss and assess, if they	Fairly familiar (n=4)	Doctoral student	Instrument

	read any difference at all, is critical to measure what you're hoping to measure here. Consider a shorter evaluation? Allow back arrow and unchecking or N/A options on all of your survey questions/pages (this is more for the Qualtrics survey, less about the CR Framework, specifically). Also some questions required response, but none of the response options were applicable (again, allow an N/A option)"			

Note: Survey respondent could have posed questions or provided perspectives during Part 1 ("Please discuss what implicit or explicit assumptions that you made about race in your research", "I'm ready to continue. I have some questions. Someone from the study will provide responses within 2 business days. Please do not complete CR Framework for the article analysis before you hear back from us." This appeared after each model of the training", "Please provide any feedback on the Critical Race Framework training" or Part 2 ("Please provide any additional feedback on the Critical Race Framework".)

[1]The number denotes how each rater self-rated in familiarity for each construct. All raters self-rated the same value; hence, n=4.

3. Results of Phase II

The study conducted 132 screenings for eligibility, which included duplicate users. Forty-three distinct participants enrolled in the study. We retained the most complete response from duplicate participants (n=6). Seven enrollees were excluded because of non-response following study enrollment.

Descriptive Statistics

All respondents and complete cases shared characteristics. More than 60% held a Ph.D. (64% vs. 64%), was employed at a college or university (69% vs. 68%), and self-identified as women (76% vs. 68%), respectively (Table 27). The mode for the number of years with terminal degree in public health was 1-5 years in each group. Over half of the sample (n=12, 55%) resided in the US South. Differences between all respondents and complete cases were not significant. As expected in a 2 x 2 contingency table, our degree of freedom was 1.

Table 27: Summary Statistics Phase II N=36			
	All Respondents	**Complete Cases** n=22	**P-Value[1]**
Terminal Degree Type1 (n=36)	**n (%)**	**n (%)**	
Ph.D.	23 (64)	14 (64)	1.0
Dr.P.H.	6 (17)	5 (23)	
Other[2]	3 (8)	2 ()	
D.Sc.	2 (6)	-	
Ed.D.	-	1 (5)	
Years with terminal degree in public health1	**n (%)**	**n (%)**	
<1	2 (6)	2 (9)	
1-5	13 (36)	6 (27)	
6-10	8 (22)	5 (23)	
11-15	4 (11)	4 (18)	
16-20	4 (11)	1 (5)	
21-25	3 (8)	2 (9)	
26-30	1 (3)	1 (5)	
31-40	0	0	
41-45	1 (3)	1 (5)	
>45	0	0	
Employer Type (n=36)	**n (%)**	**n (%)**	
College/university	25 (69)	15 (68)	1.0
Hospital/health system	6 (17)	3 (14)	
Other[3]	3 (8)	2 (9)	
Nonprofit (other than higher education and hospital)	2 (6)	2 (9)	
Government Employer	**n (%)**	**n (%)**	
No	25(69)	15 (68)	
Yes	11 (31)	7 (32)	
Primary Role (n=36)	**n (%)**	**n (%)**	
Teaching faculty	13 (36)	6 (27)	.471
Non-teaching faculty	12 (33)	8 (36)	
Other	6 (17)	5 (23)	
Researcher (outside of academia)	3 (8)	1 (5)	
Non-research administrator (outside of academia)	2 (6)	2 (9)	
US Region	**n (%)**	**n (%)**	
Northeast	11 (31)	6 (27)	
Midwest	5 (14)	2 (9)	
South	16 (44)	12 (55)	
West	4 (11)	2 (9)	
Public Health Specialty (n=36)	**n (%)**	**n (%)**	
Behavioral Science and Health Education	36 (100)	7 (32)	1.0
Epidemiology	28 (78)	10 (45)	0.32
Community Health	22 (61)	8 (36)	
Environmental Health	22 (61)	5 (23)	
Biostatistics	15 (42)	2 (9)	
Other[4]	14 (39)	3 (14)	
Maternal and Child Health	10 (28)	3 (14)	
Health Policy and Management	6 (17)	4 (18)	
Health Promotion and Education	6 (17)	3 (14)	
Health Administration	3 (8)	1 (5)	
Nutrition	3 (8)	3 (14)	

Global Health	2 (6)	0	
Public Health Policy and Management	2 (6)	0	
International Public Health	1 (3)	0	
Gender	**n (%)**	**n (%)**	
Woman	26 (76)	15 (68)	
Man	6 (18)	5 (23)	
Prefer not to say/Non-binary/none of above	2 (6)	2 (9)	

*If significant at the 0.05 level (2-tailed).

[1]We dichotomized variable (e.g., Ph.D. versus all other degree types) to conduct significance testing using Pearson chi-square when 0 cells have an expected count less than five (i.e., PhD, epidemiology) and Fisher's exact text when at least 1 cell has an expected cell count less than five (i.e., college/university, teaching faculty, behavioral).
[2]Other types of terminal degrees included DNP, Sc.D., (MD, PhD).
[3]Government or retired.
[4]disability, primary prevention, occupational health, social epidemiology, population health, SOGI demographics, dental public health, demography

Table 28 provides results on publication attributes and research practices, and agreement with study premise and federal race definitions. Nearly one-third (n=13, 36%) had more than 90% of published research that used a racial variable. This percentage increased in Phase II (n=9, 41%). Most respondents in Phase I (24, 67%) and Phase II (n=16, 73%) had more than half of their research to include a variable for race.

Most respondents never or rarely provided a definition of race to study participants: 60% (all) and 59% (complete cases). Sometimes or often collapsing racial groups in analysis was most common: 80% (all) and 87% (complete cases). Most respondents often or always assumed that racial identification is often or always stable: 60% (all) and 61% (complete cases). There were similar distributions on whether respondents sometimes, often, or always considered potential threats to statistical inferences due to inclusion of a variable for race.

Agreement with the study premise had similar distribution among the partial and complete cases. The most common response was "agree". The federal definition of race exerted major influence ("important influence" or "very important influence") on respondents' practices

in data collection and analysis: 65% (partial) and 64% (complete). “Negative” and “neutral” perceptions of the federal definition had similar distributions.

Table 28: Demographic Data for Full and Partial Respondents		
	All Respondents	**Complete Cases** n=22
% of Published Research with Race Variable (n=36)	**n (%)**	**n (%)**
None	1 (3)	1 (5)
<11%	4 (11)	2 (9)
11-25%	4 (11)	2 (9)
26-50%	3 (8)	1 (5)
51-75%	6 (6)	4 (18)
76-90%	5 (10)	3 (14)
>90%	13 (36)	9 (41)
Research Practices (n=35)		
Provide a definition of race to study enrollees	**n (%)**	**n (%)**
Never	13 (37)	7 (32)
Rarely	8 (23)	6 (27)
Sometimes	8 (23)	5 (23)
Often	2 (6)	1 (5)
Always	4 (11)	3 (14)
Collapse racial groups in analysis	**n (%)**	**n (%)**
Never	2 (6)	1 (5)
Rarely	3 (9)	1 (5)
Sometimes	13 (37)	7 (32)
Often	15 (43)	12 (55)
Always	2 (6)	1 (5)
Assume that race is stable for individuals in your research	**n (%)**	**n (%)**
Never	6 (17)	3 (14)
Rarely	4 (11)	2 (9)
Sometimes	4 (11)	3 (14)
Often	9 (26)	8 (36)
Always	12 (34)	6 (27)
Consider potential threats to statistical inferences due to inclusion of a variable for race	**n (%)**	**n (%)**
Never	2 (6)	2 (9)
Rarely	4 (11)	2 (9)
Sometimes	9 (26)	6 (27)
Often	11 (31)	8 (36)
Always	9 (26)	4 (18)
Agreement with Study Premise (n=34)	**n (%)**	**n (%)**
Strongly disagree	3 (9)	0
Disagree	2 (6)	2 (9)
Neutral	6 (18)	3 (14)
Agree	15 (44)	11 (50)
Strongly Agree	8 (24)	6 (27)
Influence of OPM race definitions on race data collection and analysis practices? (n=34)	**n (%)**	**n (%)**
No influence	2 (6)	1 (5)
Limited influence	1 (3)	0
Moderate influence	9 (26)	7 (32)
Important influence	14 (41)	9 (41)

Very important influence	8 (24)	5 (23)
Opinion about OPM race definitions	**n (%)**	**n (%)**
Very negative	3 (9)	1 (5)
Negative	13 (38)	9 (41)
Neutral	13 (38)	8 (36)
Positive	3 (9)	2 (9)
Very positive	2 (6)	2 (9)
Note:		

Missingness Analysis

The analytic sample included thirty-six respondents (n=36). The missingness rate was 26% (618 out of 2376 data points). Twenty-one respondents (n=21, 58%) completed the entire survey. Fifteen respondents dropped out prior to survey completion (n=15, 42%). The attrition was 5.5% (2 out of 36) through the question prior to the introduction of the Critical Race Framework, a page containing a video-based training, transcript, and open-ended survey question. Eleven or 73% (11 out of 15) of all survey dropouts exited the study after answering an open-ended question on social identity. This group of eleven had completed approximately 25 questions or 38% of all survey questions. Qualtrics user data indicated that the Critical Race Framework page was not displayed, meaning that study participants provided a response on social identity and did not click on the button to proceed. Thus, the Critical Race Framework cannot explain dropout.

There were no significant findings based on results from five tests statistical tests. A possible scenario is that the data could be MAR because those who dropped out had lower thresholds for survey fatigue but would not have differed otherwise in perceptions of the Critical Race Framework than the complete sample. However, survey fatigue was not directly assessed. We cannot conclude that the data were MAR based on survey fatigue or other observable factors.

We determined that the data were MNAR, as an unknown non-observable factor explained attrition. We cannot say with any degree of certainty that the survey itself affected

attrition. Unlike Phase 1, which occurred during a busy time at the end of the academic year, Phase II was open over a longer period (August 2023 – January 2024). Timing was a less plausible explanation during Phase II. Prompted by data results, the PI met with some members of the dissertation committee in late December 2023 to discuss next steps. Considering study aims to conduct exploratory factor analysis (EFA), we discussed a possible alternative involving imputation of values for up to 22 samples (all incomplete cases that enrolled in the study) to perform EFA. Imputing missing data with MNAR is not generally a valid method because it is unlikely to produce unbiased estimates (Bhaskaran & Smeeth, 2014). Alternatively stated, imputation seeks to reduce biased estimates, advisable only to the extent that factors associated with missingness are observable and can reduce bias. In the context of this study, imputation has no clear advantage other than to increase the sample size for EFA.

Considering MNAR, imputation raised major concerns of statistical and practical significance. First, we did not have a scientific or theoretical basis to have reasonably assumed that imputation would attenuate bias. In fact, it may lead to the opposite outcome – loss of power and biased estimates. Second, EFA was the final analysis in Phase I. We cannot be certain that imputed data are more adequate than complete case analysis to overcome the minimum sample size threshold, if that was not adequate. A fuller discussion on sample size adequacy appears in Chapter 4. Third, imputation would involve replacing missing values on acceptability, appropriateness, feasibility, and satisfaction for 14 (including partial respondents) to 22 (including all study enrollees) samples. Most never saw the Critical Race Framework training page based on user data analytics. The framework served as the critical research question. Consequently, we did not impute data for Phase II.

RQ1. To what extent is the Critical Race Framework 2.0 web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to research constructs? (Phase II).

Table 29 displays data results on acceptability. Sixty percent or more (≥60%) agreed that the Critical Race Framework was acceptable for each scale item, except "meets approval" (n=10, 45%). Less than 10% disagreed with acceptability, except "no objection" that garnered 18% (n=4) of "completely disagree" or "disagree" responses.

Table 29: Phase II Acceptability

	Completely disagree/disagree n (%)	**Neither agree nor disagree** n (%)	**Completely Agree/Agree** n (%)
Meets approval	2 (9)	10 (45)	10 (45)
Appealing	2 (9)	4 (18)	16 (73)
"I welcome"	2 (9)	2 (9)	18 (81)
No objection	4 (18)	2 (9)	16 (73)
"I like"	2 (9)	5 (23)	15 (68)

N=22

Note: Scale questions appeared as follows: "Critical Race Framework meeting my approval," "Critical Race Framework is appealing to me," "I welcome the Critical Race Framework," "I have no objection to the Critical Race Framework," and "I like the Critical Race Framework." The Likert scale included options for completely disagree, disagree, neither agree nor disagree, agree, and completely agree.

Satisfaction ("satisfied" or "strongly satisfied") was the most common response for perceptions on the training and instrument by a majority (Table 30).

Table 30: Phase II Satisfaction

	Dissatisfied/Strongly Dissatisfied	**Neutral**	**Satisfied/Strongly Satisfied**
Training	3 (14)	7 (32)	12 (55)
Instrument	4 (18)	4 (18)	14 (64)
	Mean (SD)		
Satisfaction Index	7.0 (1.3)		

Note: Satisfaction was assessed using a single-item Likert scale for two questions on satisfaction, "Please rate your satisfaction with the Critical Race Framework training" and "Please rate your satisfaction with the Critical Race Framework tool". The options were strongly disagree, disagree, neutral, agree, strongly agree. Collapsed data were reported (strongly disagree/disagree, neutral, agree/strongly agree) A summation index using these two questions was created for analysis. The maximum value for satisfaction is 10.

Approximately one-in-six respondents was dissatisfied (“dissatisfied” or “strongly dissatisfied”). The mean Satisfaction Index score was 7.0 out of a maximum score of 10.

More than 50% of the respondents found the CR Framework to be feasible (“completely agree” or “agree”) as in implementation (n=12, 55%), possible (n=16, 73%), doable (n=15, 68%), and workable (n=15, 68%)(Table 31). Nearly half found it easy to use (n=10, 45%). Roughly, one-third of respondents were neutral about the feasibility of the CR Framework.

Table 31: Phase II Feasibility (n=22)

Scale Item	Completely disagree/disagree	Neither agree nor disagree	Completely Agree/Agree
Implementable	3 (14)	7 (32)	12 (55)
Possible	1 (5)	5 (23)	16 (73)
Doable	1 (5)	6 (27)	15 (68)
Easy to use	5 (23)	7 (32)	10 (45)
Workable	1 (5)	6 (27)	15 (68)

Note: Scale questions appeared as follows: "The Critical Race Framework seems implementable," "The Critical Race Framework seems possible," "The Critical Race Framework seems doable," "The Critical Race Framework seems easy to use," and "The Critical Race Framework seems workable." The Likert scale included the following options: completely disagree, disagree, neither agree nor disagree, agree, and completely agree.

Sixty percent or more of respondents agreed (“completely agree” or “agreed) that the CR Framework was appropriate based on three measures: fitting (64%), suitable (64%), and applicable (68%)(Table 32). Fifty-nine percent were neutral about the CR Framework being a “good match”. Respondents were split between neutral and agree responses for “well-aligned,” respectively 41% and 45%.

Table 32: Phase II Appropriateness
(n=22)

Scale Item	Completely disagree/disagree	Neither agree nor disagree	Completely Agree/Agree
Fitting	2 (9)	6 (27)	14 (64)
Suitable	2 (9)	6 (27)	14 (64)
Applicable	1 (5)	6 (27)	15 (68)
Good match	2 (9)	13 (59)	7 (32)
Well-aligned	3 (14)	9 (41)	10 (45)

Note: Scale questions appeared as follows: "The Critical Race Framework seems fitting," "The Critical Race Framework seems suitable," "The Critical Race Framework seems applicable," "The Critical Race Framework seems like a good match," and "The Critical Race Framework seem well aligned." The Likert scale included the following options: completely disagree, disagree, neither agree nor disagree, agree, and completely agree.

Audience and Setting Appropriateness

More than three-fourths of respondents felt that the Critical Race Framework was appropriate for public health faculty (91%), master's students (82%), public health researchers (82%), doctoral students (77%), and postdoctoral fellows (77%)(Table 33). A majority felt that the Critical Race Framework was appropriate for faculty development (73%), peer-reviewed journals (64%), and instructional settings (55%).

Table 33: Audience and Setting Appropriateness

Audiences		**Settings**	
Audience Type	**n (%)**	**Type of Settings**	**n (%)**
Public health faculty	20 (91)	Faculty development in academia	16 (73)
Master's Students	18 (82)	Peer-reviewed journals	14 (64)
Public health researchers	18 (82)	Classroom/Instruction	12 (55)
Doctoral Students	17 (77)	Professional development	10 (45)
Postdoctoral fellows	17 (77)	Policymaking	8 (36)
Undergraduates	9 (41)	Public health planning/implementation	8 (36)
		Evaluating funding requests	8 (36)
		Consulting	5 (23)

n=22
Note: The question provided a list from which study participants could any option, "Please select the type of learner/employee and settings for which you feel that Critical Race Framework is appropriate."

We identified three *a priori* themes based on open-response comments: study design, instrumentation, and training. Additional themes emerged: overall, impact on field and practice, and study premise (Table 34). All qualitative data were analyzed and organized. A single possible reliability threat emerged from the *measures of fit* scale among two respondents: "I had difficulty completing the feasibility and appropriateness scales because no context was provided to ground my responses" and "Some of the questions about the project are unclear. For example, under appropriateness, the framework seems like a good match...for what? Aligned...with what?" It appeared that both respondents would have liked a more direction on these scales.

Table 34: Phase II Thematic Analysis	
Themes	**Qualitative Data**
General	"This framework seems doable and feasible." "I appreciate the tool and the video, I only marked down because it still feels rather abstract and complicated to me" "Overall, I appreciate the effort of this project to improve how data on race are collected, analyzed, and interpreted.
Study Design	"I had difficulty completing the feasibility and appropriateness scales because no context was provided to ground my responses." "Some of the questions about the project are unclear. For example, under appropriateness, the framework seems like a good match...for what? Aligned...with what? I was also left wondering about the application of this framework."
Instrument	"I think people may have difficult[y] determining what is low, moderate, high quality. I know you provided definitions, but I think seeing examples would help people." "I was surprised at the use of the term 'critical race theory' overall as this content does not align with what I think of as the core principles of CRT."
Training	"I was thinking some examples would help, (like 'here is a hypothetical form and the limitations of it...I found old CDC NHANES data that listed 3 categories for "race": white, black, Mexican. It was appalling. Maybe as you expand the work you could show how weak race data collection instruments not only give you bad data but cause psychosocial distress for some respondents... thank you for this project! Very interesting." "It sounded like some but not all of the people in the video were lip synching, or their voice was dubbed or something. It was very distracting." "I like the framework but think it will be very hard to use without more training on the four concepts. The training video is not enough to do this. One needs more of a background and reading on internal/external validity (e.g., Campbell and Stanley, etc.) and on measurement concepts of validity and reliability."
Impact on Field and Implementation	"My only comment would be what do journals and researchers do if their research does not meet the standards laid out in the framework? Does this mean the research should not be published or does this mean that the researcher should provide a disclaimer statement?" "It's not entirely clear how this tool is intended to be implemented and utilized. How would the practice of research be changed by this tool? Is the intention for researchers to utilize this tool in a similar manner as a literature review framework, for examining existing literature's conceptualization of race? Or would investigators apply the framework for study/questionnaire design in their own research, and/or in other ways? In addition, while I see its utility for some challenges associated with race data/data collection, I'm not sure how it would influence some challenges it's proposed to address. For example, it would be great if we could collect data on all possible combinations of race, but we would still be left with very small sample sizes for particular groups/combinations that would be too small to analyze. Depending on the intended application, I could also see this being challenging and time consuming to implement in certain settings."
Support for Study Premise	"I am not saying it isn't important to talk about whether there is a "true value" for race, (I) just don't think there is one. Just as we talk about gender, there is an important distinction between how people perceive you and how you identify. Colorism plays a role Also how long someone has been in the US-- people from outside the US may not relate to the racial categories here (which don't include mestizo or Metis etc.) All this is important to capture." "(I)nternal validity is compromised if we simplify the measurement of race and ethnicity." "Also - it would appear as though an underlying assumption of the framework is that the use of racial categories in health research should be retained but viewed more critically as it impacts interpretations & conclusions."
Note: We identified three a priori themes: study design, instrumentation, and training. The survey and article critique were the study design. The instrumentation concerned the CR Framework itself. The training pertained to the web-based training video in Phase I. We sought data on the appropriate length, quality, and areas of improvement.	

RQ3.What are demographic and research associations with perceptions of the Critical Race Framework 2.0? (Phase II). *Aim 3*) To determine whether demographic and research factors are associated with perceptions of the Critical Race Framework.

Table 35 lists data results for 40 pairwise comparisons (RQ3). Correlation analysis with the Familiarity Index score and the appropriateness summation index was significant at the 0.05 alpha level, as was Skills Composite and appropriateness. Upon further investigation, these are explained by two respondents with high or very high Skills Composite and Familiarity Index scores and low perceived appropriateness (not shown).

Table 35: Significance Testing for Measures of Fit and Demographics and Research

Measures	Acceptability	Appropriateness	Feasibility	Satisfaction
1. Employer[1]	0.68	0.84	0.45	0.68
2. Primary Role[2]	1.00	0.91	0.33	0.59
3. Epi/Bio vs All[3]	0.72	0.67	0.25	0.67
4. PhD vs Others[4]	1.00	0.37	0.24	0.62
5. Yrs. PhD Held[5]	-0.03	0.03	-0.09	-0.04
6. Years in PH Research[6]	-.021	-0.11	-0.30	-0.20
7. Years in PH Teaching[7]	-0.09	0.00	-0.06	-0.02
8. Premise Agreement[8]	-0.05	-0.22	-0.20	-0.22
9. Familiarity Index[9]	-.13	-.40*	0.06	-0.20
10. Skills Composite[10]	.021	-.36*	-0.09	-0.12

Note: The Mann–Whitney *U* test statistic is displayed for analyses in measures 1-4. Kendall's tau-b (τ_b) correlation coefficients are displayed for measures 5-10.

*Correlation is significant at the 0.05 level (2-tailed).

[1]**Employer** – This variable was dichotomized as: college/university versus hospital/health system, nonprofit (other than higher education), for-profit, small business, self-employed, unemployed, and an "other" option (open response).
[2]**Primary role** – This variable was dichotomized as: teaching faculty versus non-teaching faculty, researcher (outside of academia), non-research administrator (outside of academia), and other (open response).
[3]**Epi/Bio vs All** – This variable was dichotomized as: epidemiology versus behavioral science and health education, community health, environmental health, biostatistics, maternal and child health, health policy and management, health promotion and education, health administration, nutrition, global health, public health policy and management, international public health, and other (open response).
[4] **PhD vs others** – This variable was dichotomized as: Ph.D. versus Ed.D., Dr.P.H., M.D., D.O., D.Sc., D.P.E., D.B.H., D.H.Sc., D.S.W., and other (open response).
[5]**Yrs. PhD Held** – This variable used ranges for a continuous variable (in parentheses): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12).
[6]**Years in PH Research** – This variable used ranges for a continuous variable (in parentheses): none (0), <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12).
[7]**Years in PH Teaching** – This variable used ranges for a continuous variable (in parentheses): <1 (1), 1-5 (2), 6-10 (3), 11-15 (4), 16-20 (5), 21-25 (6), 26-30 (7), 31-35 (8), 36-40 (9), 41-45 (10), 46-50 (11), 51 years or more (12)
[8]**Agreement on Premise** – Participants were asked to indicate their agreement with the research premise, "The premise for this study is that race introduces potential threats to study quality in the areas of reliability, validity, internal validity, and external validity." We collapsed this variable: disagree ("disagree and "strongly disagree"), neutral, and agree ("agree" and "strongly agree").
[9]**Familiarity Index** – Familiarity Index consisting of five-item composite score was created using each self-assessed score on familiarity for each area of critical appraisal and familiarity with statistical theory and analysis. The maximum score is 20.
[10]**Skills Composite** - Skills Composite Index was a summation of two questions, "Please rate your quantitative and qualitive research skills". It used the following Likert scale: very poor, poor, average, very good, excellent. The maximum skills score was 10.

RQ4. What is the reliability and validity evidence for the Critical Race Framework 2.0? (Phase II). Aim 4) To determine the quality of reliability and validity evidence for the Critical Race Framework.

We previously reported Phase I results on straight-lining. Non-differentiation bias was a focus of our reliability assessment for Phase II. Table 36 displays the results using the simple non-differentiation method. The range of proportions ranged from 0.0 – .67. The relevance questions on external validity had the highest proportion (0.67) with 14 respondents. The scale-level proportion for overall measure of fit was zero, meaning no indication of straight-lining. Two individuals provided the same response individually to all 20 questions on relevance.

Table 36: Simple Non-Differentiation Method by Measures of Fit and Article Evaluation

Measure	N	Values (Proportion)
Overall Measures of Fit (1-3)[1]	0	.00
1. Acceptability	7	.31
2. Feasibility	6	.27
3. Appropriateness	11	.50
Relevance (Overall)[2]	2	.10
1. Reliability	4	.19
2. Validity	10	.48
3. Internal Validity	5	.24
4. External Validity	14	.67

Note: The simple nondifferentiation measure was calculated as the percentage of identical responses.
[1]N=22
[2]N=21

We assessed reliability in internal consistency and correlation analysis. Results from measuring internal consistency are reported in Table 37. The Cronbach's α approached or exceeded high internal consistency (≈0.90).

Table 37: Measures of Internal Consistency

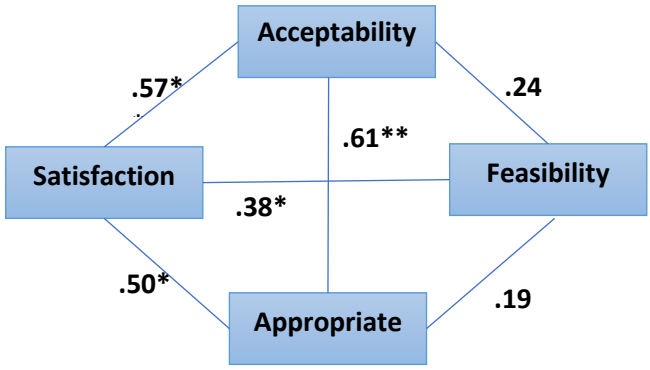
Scale	Mean (Variance)	Cronbach's α
Acceptability	3.7 (0.05)	.91
Feasibility	3.6 (0.04)	.87
Appropriateness	3.5 (0.03)	.91

N=22
Note: The *acceptability* scale questions appeared as follows: "Critical Race Framework meeting my approval," "Critical Race Framework is appealing to me," "I welcome the Critical Race Framework," "I have no objection to the Critical Race Framework," and "I like the Critical Race Framework." The Likert scale included options for completely disagree, disagree, neither agree nor disagree, agree, and completely agree. The *feasibility* scale questions appeared as follows: "The Critical

Race Framework seems implementable," "The Critical Race Framework seems possible," "The Critical Race Framework seems doable," "The Critical Race Framework seems easy to use," and "The Critical Race Framework seems workable." The Likert scale included the following options: completely disagree, disagree, neither agree nor disagree, agree, and completely agree. The *appropriateness* scale questions appeared as follows: "The Critical Race Framework seems fitting," "The Critical Race Framework seems suitable," "The Critical Race Framework seems applicable," "The Critical Race Framework seems like a good match," and "The Critical Race Framework seem well aligned." The Likert scale included the following options: completely disagree, disagree, neither agree nor disagree, agree, and completely agree.

Figure 5 displays results of correlation analysis for four measures of fit. Kendall's tau-b (τb) correlation coefficients are displayed. Four of six pairwise correlations were statistically significant. Satisfaction accounted for three of these. The strength of correlation was greatest (p<0.001) for 1) satisfaction and acceptability and 2) acceptability and appropriateness. Satisfaction was significantly correlated with feasibility and appropriateness. Pairwise comparisons for 1) acceptability and feasibility and 2) feasibility and appropriateness were not statistically significant.

Figure 5: Bivariate Correlations for Measures of Fit



N=22
Note: Kendall's tau-b (τ_b) correlation coefficients are displayed.
*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

Table 38 displays the content validity indexes (CVI) and Kappa designating agreement of relevance (k*). Each I-CVI exceeded ≥0.75 (excellent), except for “existence of a “true value(s)” for race” (I-CVI=0.70). All S-CVIs exceed 0.75: reliability (0.80), validity (0.89), internal validity (0.89), and external validity (0.95). S-CVIs were excellent in content validity. Each item

in the CR Framework had an excellent Kappa value (k*≥0.75), except item 4 ("Existence of a "true value(s)" for race") that appeared in the reliability section (k*=0.69).

Table 38 - Content Validity Index (CVI) and the Kappa Designating Agreement of Relevance (k*) of the Critical Race Framework

	CVI	Interpretation[1]	k*	Interpretation[2]
Reliability				
1. Reliability evidence of survey tool(s) used to collect racial identity	0.90	Excellent	0.90	Excellent
2. Potential participant sources of measurement error in race data collection?	0.75	Excellent	0.75	Excellent
3. Potential sources of measurement error due to the race data collection tool(s)	0.90	Excellent	0.90	Excellent
4. Existence of a "true value(s)" for race	0.70	Not Excellent	0.69	Good
Scale	0.80	Excellent	—	—
Validity				
5. Construct or meaning of race used in study	0.90	Excellent	0.90	Excellent
6. The inclusion of multiracial identity to construct or meaning of race used in study	0.90	Excellent	0.90	Excellent
7. Characteristics intended to differentiate racial groups	0.85	Excellent	0.85	Excellent
8. Heterogeneity within racial groups	0.90	Excellent	0.90	Excellent
Scale	0.89	Excellent	—	—
Internal Validity				
9. Potential threats to internal validity due to quality of reliability and validity of the race variable alone	0.95	Excellent	0.95	Excellent
10. Population data estimates for all possible combinations of race based on race data collection tool(s)	0.75	Excellent	0.75	Excellent
11. Methods to provide participants with study construct or meaning of race during data collection	0.90	Excellent	0.90	Excellent
12. Data results of all possible combinations of race based on original race data collection tool(s)	0.85	Excellent	0.85	Excellent
13. Justification to combine, exclude, or change original race data reporting	0.90	Excellent	0.90	Excellent
14. Meeting statistical assumption of independence considering racial grouping	0.80	Excellent	0.80	Excellent
15. Limitations of statistical reasoning due to a race variable	1	Excellent	1	Excellent
16. Interpretability of data results on racial group analysis	1	Excellent	1	Excellent
Scale	0.89	Excellent	—	—
External Validity				
17. Limitations of external validity due to the construct or meaning of race used in study (validity)	1	Excellent	1	Excellent
18. Limitations of external validity due to analytical treatment of race	0.85	Excellent	0.85	Excellent
19. Limitations of external validity due to within-group racial heterogeneity	0.95	Excellent	0.95	Excellent

20. Limitations of external validity due to social and political changeability of race	1	Excellent	1	Excellent
Scale	0.95	Excellent	—	—

N=21

Note: We divided the number of experts with acceptable scores ("fairly relevant" or "highly relevant") by the total number of scores among raters. In CVI analysis by question (I-CVI), the denominator was the total number of raters. The denominator by construct was the number of raters multiplied by the number of questions for that construct (S-CVI). We consider an excellent CVI to be ≥.75 consistent with cutoffs previously published in the literature (Larsson et al., 2015). We calculated a modified Kappa value, using the formula: (k* = (CVI- Pc)/(1-Pc). The probability of chance occurrence (Pc) formula for a binominal random variable was used: Pc = [N!/A!(N–A)!]*5^N. *N* was the number of raters. *A* was the number agreeing on fairly or highly relevant.

[1]Excellent CVI is ≥.75
[2]Fair = k* of 0.40 to 0.59; Good = k* of 0.60 to 0.74; and Excellent = k* of 0.75 to 1.00

Exploratory Factor Analysis

We conducted four factor analyses for each construct in the CR Framework. The total sample size was 21. The total number of item responses was 410. Table 39 displays results from factorability indices. All results were non-zero determinants, indicating that there are linear combinations of factors and lack of high multicollinearity. The Bartlett's Test of Sphericity was significant in each test, meaning that the determinant was statistically different from zero. The KMO was "middling" for EFA 1 (0.70) and EFA 2 (0.73) and "mediocre" for EFA 3 (0.67) and EFA 4 (0.60).

Table 39: Factorability Results for Exploratory Factor Analyses

		Bartlett's Test of Sphericity			
	Determinant	**Chi-Square**	***df***	**Sig.**	**KMO***
EFA 1	0.151	33.7	6	<0.001	0.70
EFA 2	0.060	50.3	6	<0.001	0.73
EFA 3	0.016	68.6	28	<0.001	0.67
EFA 4	0.096	41.8	6	<0.001	0.60

Note: The table displays the output from each exploratory factor analysis: EFA 1 (reliability), EFA 2 (validity), EFA 3 (internal validity), and EFA 4 (external validity). N=21. The number of items per construct is as follows: reliability (n=4), validity (n=4), internal validity (n=8), and external validity (n=4). The number of total survey respondents equals 21.

*Kaiser-Meyer-Olkin Measure of Sampling Adequacy - marvelous (0.90 – 1.00), meritorious (0.80 – 0.89), middling (0.70 – 0.79), mediocre (0.60 – 0.69), miserable (0.50 – 0.59), and no factorability (0.00 – 0.49).

Table 40 displays results from the exploratory factor analyses. The four items for reliability (EFA 1) explained 57.6% of the total variance for the first factor. The sums of squared loadings ranged from 0.56 – 0.99. Item 3 ("Potential sources of measurement error due to the race data collection tool(s)") had the highest loading. Item 4 ("Existence of a "true value(s)" for race") had the lowest loading (0.56).

The total variable explained for validity was 69%. The sums of squared loadings approached or exceeded 0.70. "Characteristics intended to differentiate racial groups" had the lowest loading (0.68); whereas "Heterogeneity within racial groups" had the highest (0.92). The eight items for EFA 3 explained 39.6% of the total variance for the first factor. The loadings varied from 0.33 ("Methods to provide participants with study construct or meaning of race during data collection") to 0.82 (Population data estimates for all possible combinations of race based on race data collection tool(s)"). The final EFA had a total variance explained of 60%. The lowest sums of squared loading was 0.67 for item 20 ("Limitations of external validity due to social and political changeability of race"). The highest (0.85) was for item 19 ("Limitations of external validity due to within-group racial heterogeneity").

Table 40: Results of Exploratory Factor Analyses

	Sums of Squared Loadings	**Communality[1]**	**Total Variance Explained**
Reliability (EFA 1)	—	—	57.6%
1. Reliability evidence of survey tool(s) used to collect racial identity	.75	.56	—
2. Potential participant sources of measurement error in race data collection?	.68	.46	—
3. Potential sources of measurement error due to the race data collection tool(s)	.99	.97	—
4. Existence of a "true value(s)" for race	.56	.32	—
Validity (EFA 2)			69.1%
5. Construct or meaning of race used in study	.86	.74	
6. The inclusion of multiracial identity to construct or meaning of race used in study	.84	.71	
7. Characteristics intended to differentiate racial groups	.68	.46	
8. Heterogeneity within racial groups	.92	.85	

Internal Validity (EFA 3)			39.6%
9. Potential threats to internal validity due to quality of reliability and validity of the race variable alone	.72	.52	
10. Population data estimates for all possible combinations of race based on race data collection tool(s)	.82	.67	
11. Methods to provide participants with study construct or meaning of race during data collection	.33	.11	
12. Data results of all possible combinations of race based on original race data collection tool(s)	.77	.59	
13. Justification to combine, exclude, or change original race data reporting	.44	.20	
14. Meeting statistical assumption of independence considering racial grouping	.63	.40	
15. Limitations of statistical reasoning due to a race variable	.575	.33	
16. Interpretability of data results on racial group analysis	.60	.36	
External Validity (EFA 4)			60.3%
17. Limitations of external validity due to the construct or meaning of race used in study	.84	.70	
18. Limitations of external validity due to analytical treatment of race	.74	.54	
19. Limitations of external validity due to within-group racial heterogeneity	.85	.72	
20. Limitations of external validity due to social and political changeability of race	.67	.45	
Note: The table displays the output from each exploratory factor analysis: EFA 1 (reliability), EFA 2 (validity), EFA 3 (internal validity), and EFA 4 (external validity). N=21. The number of items per construct is as follows: reliability (n=4), validity (n=4), internal validity (n=8), and external validity (n=4). [1]The extracted communalities are displayed.			

4. Results of Phase III

Table 41 displays the results of the structured article appraisal using the Critical Race Framework by rater. Rater 1 and Rater 2 evaluated the same articles (n=10), as did Rater 1 and Rater 3 (n=10). Rater 1 found a large percentage of moderate- or high-quality discussion for item 16 “Interpretability of data results on racial group analysis” (n=16). Fifteen percent (15%) or less of articles had moderate or high ratings for all other items. Rater 1 did not score moderate or high for any article for 12 of the 20 items.

Rater 2 determined moderate or high quality was not common. The highest number of articles was 5 (50%) for “9. Potential threats to internal validity due to quality of reliability and validity of the race variable alone”. Rater 2 tended to score more articles as moderate or high

than Rater 1, but the number of articles for each item was low. Sixteen out of twenty item scores had two or fewer articles with moderate or high discussion.

Overall, Rater 3 did not determine that the articles scored high or moderate as evaluated against the CR Framework. The highest rated items were "Item 8 - heterogeneity within racial groups" (7 articles) and "Item 16 - Interpretability of data results on racial group analysis" (7 articles). Fifteen items did not score high or moderate for any article.

Table 41: Moderate or High Article Ratings Using Critical Race Framework

	Rater 1	**Rater 2**	**Rater 3**
CR Framework Items	n (%)	n (%)	n (%)
1. Reliability evidence of survey tool(s) used…	0	1 (10)	0
2. Potential participant sources of measurement…	0	1 (10)	1 (10)
3. Potential sources of measurement error…	1 (5)	3 (30)	0
4. Existence of a "true value(s)" for race	0	2 (20)	0
5. Construct or meaning of race used in study	0	2 (20)	0
6. The inclusion of multiracial identity…	0	3 (30)	0
7. Characteristics intended to differentiate…	2 (10)	2 (20)	0
8. Heterogeneity within racial groups	2 (10)	2 (20)	7 (70)
9. Potential threats to internal validity…	0	5 (50)	0
10. Population data estimates for all…	0	1 (10)	0
11. Methods to provide participants…	0	1 (10)	0
12. Data results of all possible combinations…	0	1 (10)	0
13. Justification to combine, exclude…	3 (15)	1 (10)	7 (70)
14. Meeting statistical assumption of…	0	1 (10)	0
15. Limitations of statistical reasoning…	0	2 (20)	0
16. Interpretability of data results on racial…	17 (85)	3 (30)	10 (100)
17. Limitations of external validity due to…	2 (10)	0	0
18. Limitations of external validity due to…	1 (5)	0	0
19. Limitations of external validity due to…	3 (15)	1 (10)	1 (10)
20. Limitations of external validity due to…	0	1 (10)	0

Note: Rater 1 and Rater 3 scored the same articles (n=10). Rater 2 and Rater 3 scored the same articles (n=10). Rater 1 evaluated a total of 20 articles. Rater 1 assessed articles from the health disparities literature. Rater 2 assessed behavioral health studies. Raters evaluated using a scale (high quality discussion, moderate quality discussion, low quality discussion, and no discussion)

Table 42 displays interrater agreement and analysis for pairwise rater comparisons. For each CR Framework item, there were 10 pairwise comparisons. Raters 1 and 2 had 70% or more absolute agreement for nine out of 20 items.

Table 42: Pairwise Interrater Agreement and Analysis				
	Raters 1, 2		Raters 1, 3	
	Match (%)	Weighted K	Match (%)	Weighted K
1. Reliability evidence of survey tool(s) used to collect racial identity	7 (70)	.52*	7 (70)	–
2. Potential participant sources of measurement error in race data collection?	8 (80)	–	5 (50)	.29
3. Potential sources of measurement error due to the race data collection tool(s)	6 (60)	-0.18	9 (90)	.62*
4. Existence of a " true value(s)" for race	7 (70)	–	9 (90)	–
5. Construct or meaning of race used in study	6 (60)	0.67	6 (60)	.2
6. The inclusion of multiracial identity to construct or meaning of race used in study	3 (30)	0.03	6 (60)	.2
7. Characteristics intended to differentiate racial groups	6 (60)	0.36	6 (60)	-.52*
8. Heterogeneity within racial groups	5 (50)	0.58	1 (10)	.06
9. Potential threats to internal validity due to quality of reliability and validity of the race	5 (50)	0.11	8 (80)	.41
10. Population data estimates for all possible combinations of race based on race data	10 (100)	–	8 (80)	-.11
11. Methods to provide participants with study construct or meaning of race during data	8 (80)	–	9 (90)	–
12. Data results of all possible combinations of race based on original race data	6 (60)	–	9 (90)	–
13. Justification to combine, exclude, or change original race data reporting	7 (70)	.76*	2 (20)	.19
14. Meeting statistical assumption of independence considering racial grouping	7 (70)	–	10 (100)	–
15. Limitations of statistical reasoning due to a race variable	6 (60)	–	9 (90)	–
16. Interpretability of data results on racial group analysis	3 (30)	.48*	1 (10)	–
17. Limitations of external validity due to the construct or meaning of race used in study	5 (50)	0.29	7 (70)	-.11
18. Limitations of external validity due to analytical treatment of race	8 (80)	0.52	5 (50)	-.05
19. Limitations of external validity due to within-group racial heterogeneity	5 (50)	-0.25	6 (60)	.19
20. Limitations of external validity due to social and political changeability of race	7 (70)	–	10 (100)	–

Note: Match is determined based on identical selection. Raters evaluated using a scale (high quality discussion, moderate quality discussion, low quality discussion, and no discussion). No weighed Kappa computed if all item ratings are a constant.

Table 43: Pairwise Interrater Agreement and Analysis for Health Disparities Articles (Dichotomized)		
	Raters 1,2	Raters 1,3
	Match (%)	Match (%)
1. Reliability evidence of survey tool(s) used to collect racial identity	9 (90)	10 (100)
2. Potential participant sources of measurement error in race data collection?	9 (90)	9 (90)
3. Potential sources of measurement error due to the race data collection tool(s)	6 (60)	10 (100)
4. Existence of a " true value(s)" for race	8 (80)	10 (100)
5. Construct or meaning of race used in study	8 (80)	10 (100)
6. The inclusion of multiracial identity to construct or meaning of race used in study	7 (70)	10 (100)
7. Characteristics intended to differentiate racial groups	7 (70)	9 (90)
8. Heterogeneity within racial groups	8 (80)	3 (30)
9. Potential threats to internal validity due to quality of reliability and validity of the race	5 (60)	10 (100)
10. Population data estimates for all possible combinations of race based on race data	9 (90)	10 (100)
11. Methods to provide participants with study construct or meaning of race during data	9 (90)	10 (100)
12. Data results of all possible combinations of race based on original race data	9 (90)	10 (100)
13. Justification to combine, exclude, or change original race data reporting	8 (80)	3 (30)
14. Meeting statistical assumption of independence considering racial grouping	9 (90)	10 (100)
15. Limitations of statistical reasoning due to a race variable	8 (80)	10 (100)
16. Interpretability of data results on racial group analysis	6 (60)	10 (100)
17. Limitations of external validity due to the construct or meaning of race used in study	9 (90)	9 (90)
18. Limitations of external validity due to analytical treatment of race	10 (100)	9 (90)
19. Limitations of external validity due to within-group racial heterogeneity	7 (70)	8 (80)
20. Limitations of external validity due to social and political changeability of race	9 (90)	10 (100)

Raters 1 and 3 had 70% or more absolute agreement for 11 out of 20 items. Table 43 displays the same data except that the scale is dichotomized (high/moderate versus low/no).

Table 44 displays results the upper-bound of the Critical Race Framework ratings for each article. This table displays the sum of "high" or "moderate" quality ratings across the Critical Race (CR)

Framework if at least one rater gave it either score. If the two raters scored the same for "high" or "moderate," we only count this once. The denominator (n=20) was used for all percentages – the number of items in the CR Framework. Rater 1 (PI) and 2 (Dyer) scored the health disparities articles (n=10). Rater 1 (PI) and Rater 3 (Woodward) scored the behavioral health articles (n=10). Fifteen articles had equal to or less than 25% "high" or "moderate" ratings across the CR Framework. Scores clusters, meaning five or more articles with "high or moderate" for either or both raters were items (by pairwise comparison): Raters 1,2 [9. Potential threats to internal validity due to quality of reliability and validity of the race variable alone (n=5), 16. Interpretability of data results on racial group analysis (n=7)] and Raters 1,3 [8. Heterogeneity within racial groups (n=7), 13. Justification to combine, exclude, or change original race data reporting (n=7), 16. Interpretability of data results on racial group analysis (n=10) (not shown).

Table 44: Upper-Bound "High" or "Moderate" Quality Scores by Article Using the 20-Item Critical Race Framework

Health Disparities Articles	n (%)	Behavioral Health Articles	n (%)
Arias et al (2002)	7 (35)	Bell et al (2018)	2 (10)
Ashing-Giwa et al (2004)	4 (20)	Cuevas et al (2020)	3 (15)
Banks et al (2006)	3 (15)	Devaraj et al (2020)	2 (10)
De Hert et al (2011)	0 (0)	Haq et al (2020)	6 (30)
Meissner et al (2006)	0 (0)	Horwitz et al (2018)	5 (25)
Mensah et al (2005)	1 (5)	Luo et al (2022)	1 (5)
Shavers (2002)	7 (35)	McClendon et al (2021)	3 (15)
Skinner (2003)	8 (40)	Parcha et al (2020)	3 (15)
Trivedi et al (2005)	15 (75)	Tuthill et al (2020).	2 (10)
Van Doorslaer (2004)	0 (0)	Whitaker et al (2018)	3 (15)

Note: This table displays the sum of "high" or "moderate" quality ratings across the Critical Race (CR) Framework if at least one rater gave it either score. If the two raters scored the same for "high" or "moderate," we only count this once. The denominator (n=20) was used for all percentages – the number of items in the CR Framework.

Chapter 5: Discussion

1. Introduction to Chapter 5

The Critical Race Framework study directly challenged long-standing research norms and practices on racial conceptualization, collection, and analysis. The common use of race is attributed to research norms rather than scientific rigor (Lin & Kelsey, 2000; Witzig, 1996). Race lacks conceptual clarity and additional key features consistent with scientific rigor (Martinez et al., 2022). The Critical Race Framework study aimed to fill a major and systematic gap in the literature to account for the inherent bias that race conceptualization, collection, and analysis presents in public health research. In our literature review, we showed that seminal texts, many critical appraisal tools, and BRFSS do not consider the effects of race on reliability, validity, internal validity, and external validity. Notable researchers in minority health have noted the conceptual and methodological issues with race yet have insisted on and supported its inclusion in disparities research (C. P. Jones, 2001; T. LaVeist et al., 2011; T. A. LaVeist, 1994).

The only similar tool – the CARMeL tool – is found in the medical education literature (Garvey et al., 2022). However, the CARMeL tool has major drawbacks by largely supporting current practices in the use of the race in research, if biological assumptions are avoided. The authors support a sociopolitical framework for race, but do not further a strong validity basis to define these differences and defend them as relevant and meaningful for science.

Phase I pilot-tested an innovation – the Critical Race (CR) Framework – using quantitative and qualitative data-gathering and identified improvements in *measures of fit* (RQ1), study design instrumentation, and training (RQ2). In Phase II, we gathered reliability (RQ4) and validity (RQ5) evidence against which to measure the quality of our critical appraisal tool. We also needed to assess potential confounding in our study (RQ3). Our research methodologies

were directly aligned with our research aim to develop an acceptable, appropriate, relevant, and feasible bias tool to structure qualitative critical appraisal and that provided evidence of reliability and validity.

Our study findings contribute to the literature. First, the Critical Race Framework is the first critical appraisal tool in public health to address the singular threat that racial taxonomy poses to data integrity. Second, we provide reliability and validity evidence to spur future research. The subsequent discussion addresses each of our six research questions for this study. We summarize the findings to our research questions in Table 45.

Table 45: Summary of Study Findings to Research Questions

Research Questions	Overall Findings
RQ1. To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs?	*Phase I:* Inconclusive, but promising if improved study design, instrumentation, and training. *Phase II:* Study findings indicate highly positive outcomes for measures of fit. We determined that the Critical Race Framework is acceptable, feasible, appropriateness, satisfactory, and that items are relevant to the related construct. However, a major caveat is that they did not apply the framework.
RQ2: What were needed improvements to the Critical Race Framework study design, instrument, and training?	Despite a small sample size, we concluded that significant revisions were required in study design, instrumentation, and training.
RQ3. What are demographic and research associations with perceptions of the Critical Race Framework?	Our data did not generally show statistically significant associations with demographic/research factors and measures of fit. Due to a small sample size, we cannot generalize our study findings beyond our study population.
RQ4. What is the a) reliability and b) validity evidence for the Critical Race Framework?	Phase II: b) Validity – Overall data results showed excellent content validity. Our results were inconclusive on internal validity and external validity. Evidence of construct validity for reliability and validity items in the CR Framework was poor to fair yet promising. We consider our results as exploratory, but a contribution to the literature nonetheless because it provides preliminary data against which future research can confirm or reject. Phase III: a) Reliability – Interrater agreement was moderate to high among rater pairs, which does not fully establish interrater reliability, but shows promise. Due to lack of confidence in significance testing, we found that interrater reliability results were inconclusive.
RQ5. What is the quality of 1) highly cited health disparities studies and 2) behavioral health studies based on the Critical Race Framework?	Twenty studies drawn from the health disparities and behavioral health literature showed low quality or no discussion when the Critical Race Framework was used.

2. Future Directions

In Chapter 1, we discussed implications of the CR Framework study for public health and behavioral health in three domains – research, practice, and policy. While the CR Framework training and tool remain under development, our study results are promising. These results can and should immediately initiate discussions and critical thought on current research practices and norms. This study delineated a realistic scope for what can be accomplished within our two-year time constraints. As such, we did not offer guidance for how an individual researcher or educator should apply this study in research development or contexts, Chapter 1 and Chapter 2 offered sufficiently rich discussion for educational enrichment and practical significance.

We cannot be sure of the pace in future improvement and full development of the CR Framework since no funding support currently exists. Although we intend to publish our study findings in a peer-reviewed journal, several future directions can accomplish the post-study aims, as outlined in Steps 15 through 19 of the Conceptual Framework for Critical Race Framework Development (Figure 2). We discuss these in the following sections – increasing sample sizes to establish reliable results and to conduct testing, assessing influences of training and demographic factors on perceptions of the CR Framework, identifying barriers to effective instruction and use, encouraging test-retest reliability and further validity testing of the CR Framework and measures of fit, and improving recruitment strategies to incent a large number of article evaluations that would be required for robust testing.

Beyond the quality testing of the CR Framework, a consensus conference or national workgroup of public health experts would be appropriate to deliberate on optimal research practices, including analytical decision-making, to overcome the limitations of datasets that only collect monolithic racial and ethnic taxonomy. Such a group can broaden Feero and colleagues'

recommendations beyond editors (Feero et al., 2024). This may mean that governmental collection efforts that are critical for public health monitoring and disparities research need to collect additional information such as lineage, cultural affiliation, assimilation, religious affiliation, historical oppressed and vulnerable populations (e.g., descendants of US enslaved and Jim Crow families (ADOS), identity changes across the life course, intergenerational poverty, combined with data on the public health economy.

US ethnic groups covered by the minimum federal standards are largely based on a notion of national heritage (e.g., Chinese, Jamaican, Mexican, Lebanese, Native Hawaiian, German), despite that countries can have much ethnic, linguistic, cultural, and religious diversity that becomes disadvantaging to minority groups within those countries are regarded monolithically. For example, "The majority of China consists of the Han people (93.3%), and 55 official minority nationalities (6.7%), most of which have their own languages, are found predominantly in the peripheral regions" (Chu et al., 1998).

Racial taxonomy disadvantages US populations such as ADOS. It is a position of this study that public health research is needed to understand individual and community health intergenerational effects due to forms of government-backed oppression and victimization such as slavery and Jim Crow. It can support efforts for ADOS reparations.

---

3. Research Question 1

To what extent is the Critical Race Framework web-based training and bias tool acceptable, satisfactory, feasible, appropriate, and relevant to a priori constructs?

Finding 1 (Phase I): *Inconclusive, but promising if improved study design, instrumentation, and training.*

Phase I was a single-site pilot study designed to assess *measures of fit* that included acceptability, feasibility, appropriateness, satisfaction, and relevance and to test a study design using a single article critique. Of the approximately 280 faculty and doctoral students who received a solicitation request, sixteen enrolled in the study, resulting in a roughly 5% response rate. Most enrollees (10 out of 16) did not answer all questions. Recruitment for this phase began near the end of the academic calendar. We immediately initiated recruitment following IRB determination. We speculated that lack of time availability adversely impacted enrollment and completion. However, we did not follow up with partial respondents – a weakness in our study – and cannot establish this claim with full confidence.

As we discuss later, we saw similar low enrollment and high attrition rates in Phase II. We speculated that there may also be unique characteristics about our study population. In published national surveys of public health researchers, we did not find that our study population was uniquely challenged to survey research. Brownson and colleagues reported a 55% responses rate among a recruitment sample of 288 public health researchers (Brownson et al., 2013). Harris and colleagues found a high completion rate (75%) among study enrollees (n=278) who were recruited from the Applied Public Health Statistics section of the American Public Health Association (Harris et al., 2018). We speculated on alternate explanations.

The topic of the study may not have been perceived as highly relevant to the sampling frame – public health researchers and educators. One comment raised this point, "(T)he full blown use of the tool will likely be limited to individuals seeking to study the use of race in research itself" (Table 26). It is possible that researchers whose primary research does not concern racial health disparities perceived low relevance to their research practices that might explain low study enrollment. We have posited, however, that the common use of race in public

health research made our study disciplinarily germane. In Chapter 2, we exposed flaws in BRFSS research methodologies, except BRFSS is illustrative of a systematic issue in public health research, as discussed in our literature review and the section on the US census. Future research should explore the norms, attitudes, and practices that may reduce low perceived relevance of the Critical Race Framework to public health education and research.

Due to MNAR, high attrition (63%), and low sample size (n=6), our data on the outcomes of interest cannot be generalized beyond the small analytic sample, even generalizability to the sample if limited due to concerns raised in our discussion. We could not determine with confidence the quality of our six *measures of fit*, consistent with the conceptual framework, given these drawbacks. Additional threats to the reliability of data were found in quantitative and qualitative analyses due to some respondents' difficulty with interpreting certain questions ("difficulty understanding") and concerns with the study design (" I got lost"), the instrument ("Be careful of double-barreled items"), and training ("The training was too long").

Finding 2: *Study findings indicate highly positive outcomes for measures of fit. We determined that the Critical Race is acceptable, feasible, appropriateness, satisfactory, and that items are relevant to the related construct. However, a major caveat is that Phase II participants did not apply the CR Framework in article critique.*

Overall, we concluded that the Critical Race Framework is acceptable. Four of the five items garnered more than 60% agreement: appealing (n=16, 73%), "I welcome" (n=18, 81%), non-objectionable (n=16, 72%), and "I like" (n=15, 68%). The single exception being "meets approval" (n=10, 45%). Nearly half were neutral on this item (n=10, 45%). Disagreement ("completely disagree" or "disagree") was uncommon for most items. Nearly, one-in-five had at least some objection ("no objection")(n=4, 18%).

A major caveat of these findings is that Phase II participants did not apply the CR Framework to an article critique. We do not know whether or how their perceptions may change. Nearly one-third was ambivalent ("neither agree nor disagree) about feasibility of the CR Framework, which is likely explained by the survey research method not accompanied by an article critique. We hypothesize the acceptability and appropriateness are less apt to change when participants can engage in using our critical appraisal tool since feasibility is focused on application. Future research is needed to assess the impact on implementation measures and the change in perceptions after training. Repeated measures analysis of variance (ANOVA) or analysis of covariance (ANCOVA) can be used to determine the training or application effect on *measures of fit*.

Our correlation analyses suggest the presence of multicollinearity among our *measures of fit*. Satisfaction was significantly correlated with acceptability (0.57), appropriateness (0.5), and feasibility (0.38). Acceptability and appropriateness were also significantly correlated (0.61). Future research should explore the model fit using structural equation modeling or regression – both would require larger sample sizes that we could attract. A weakness of our study is that we did not have a larger sample size to determine if satisfaction was an unnecessary variable in our measures of fit theory or a related variable that strengthened the model fit.

---

4. Research Question 2

What were needed improvements to the Critical Race Framework study design, instrument, and training? (Phase I).

Finding: *Despite a small sample size, we concluded that significant revisions were needed in study design, instrumentation, and training.*

Our discussion outlines key issues and changes in each of these areas prior to Phase II implementation. We relied on the conceptual development model to guide these procedures (Figure 2). Our discussion is broken down as follows: 1) study design, 2) instrumentation, and 3) training.

1. Refinement of the Critical Race Framework Study Design

Most participants spent 40-80 minutes rating and reading a single article with the CR Framework (Table 18). We estimated that the 8-10 articles that we anticipated for Phase II would require a minimum of 5.5 hours – an unrealistic research expectation given the initial lack of monetary compensation.

An aim of Phase I was to use survey responses and feedback to refine the methods for Phase II of the CR Framework study. We anticipated considerable challenges to replicate Phase I. Phase II was initially intended to provide for the same study activities as Phase I – web-based training, completing survey questions in Part 1, review of the Critical Race Framework tool, reading a research article, applying the Critical Race Framework, and completing survey questions in Part 2. However, Phase II had 8-10 individual article reviews for 30 reviewers, rather than the single review in Phase I. Phase I findings on acceptability and feasibility demonstrated that the initial Phase II plans for reviewers to apply the CR Framework 1.0 to multiple articles was not feasible.

Our study required changes to support feasibility and analysis. We considered using incentives ($100 gift cards) to encourage enrollment and completion of all study activities for Phase II as with Phase I. Funding was limited to $500, meaning that we could only offer incentives for up to five raters. This approach would produce a margin of error of 0.30 based on methods described by Gwet for calculating the percent error margin assuming five rater, three

categories, a critical value ($Z\alpha = 1.645$, $\alpha=0.10$), and five article reviews (Gwet, 2014). Even if we were successful in recruiting five highly qualified raters for five article reviews, the quality of the data would be poor and lack practical significance. In interrater reliability for this study, the number of article critiques is more important than the number of coders. Findings of low reliability in a study of the "risk of bias (ROB) instrument for non-randomized studies of exposures (ROB-NRSE)" support our conclusion on the likely poor outcomes if we pursued the original study design (Jeyaraman et al., 2020).

Thus, the decision was made to limit Phase II to web-based training and a survey. The attrition in Phase I occurred primarily between Part 1 (web-based training) and Part 2 (article review and application). It can be assumed that the article review was a driving factor associated with attrition. Thus, Phase III was an opportunity to assess interrater reliability.

2. Refinement of the Critical Race Framework

We refined the Critical Race Framework using quantitative and qualitative data. We relied on feedback from Phase I to inform the Critical Race Framework web training and tool, as discussed below. This discussion focuses on the tool itself, rather than the web-based training. A separate discussion on training appears below. The PI made significant changes to the Critical Race Framework (CR Framework 2.0) based on the data results from Phase I. Phase I was the first occasion for the Critical Race Framework to receive feedback from advanced public health researchers and educators. Survey modification occurred between the close of Phase I on June 22, 2023, and July 9, 2023. The data frequency tables and qualitive feedback in Tables 15-24 informed the CR Framework 2.0. Tables 46-47 provide a summary of changes to the Critical Race Framework 1.0. None of the questions retained their original language. Sixteen (53%) of

the questions were revised to enhance readability, limit topic scope, and reduce vocabulary load. The other questions were removed.

Table 46 – Summary of Changes to Critical Race Framework 1.0 for Phase II		
No.	**CR Framework**	**Change**
I. Reliability		
1.	Discussed reliability of racial measurement tool?	**Revised.*** Q1) *Reliability evidence of survey tool(s) used to collect racial identity*
2.	Cited prior reliability evidence?	**Deleted.†**
3.	Supported reliability of racial measurement tool with multiple measurement?	**Deleted.§**
II. Validity		
4.	Discussed what construct or meaning race intended to measure?	**Revised.*** Q5) *Construct or meaning of race used in study (validity)*
5.	Discussed relevance of racial admixture to construct or meaning of race used in study?	**Revised.*** Q6) *The inclusion of multiracial identity to the construct or meaning of race used in study (validity)*
6.	Discussed whether a "true value" exists for race?	**Revised.*** Q4) *Existence of a "true value(s)" for race*
7.	Cited validity evidence for racial measurement tool?	**Deleted.†**
8.	Discussed characteristics intended to distinguish racial groups?	**Revised.*** Q7) *Characteristics intended to differentiate racial groups (discriminant validity)*
9.	Assessed whether racial measurement tool correctly and consistently distinguished races?	**Deleted.†**
III. Internal Validity		
A. Threats to Internal Validity		
10.	Assessed threats to internal validity due to race variable? Examples include measurement error, multicollinearity, and racial differences in history, mortality, and group selection.	**Revised.*** Q9) *Potential threats to internal validity due to quality of reliability and validity of the race variable alone*
11.	Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	**Revised.*** Q15) L*imitations of statistical reasoning due to a race variable*
B. Methodology		
12.	Discussed limitations of race to establish quality of study design? Examples include data limitations to compare to known population (e.g., lack of racial admixture data), establishing success of random assignment, differences in racial data collection/tools with sample and study populations.	**Deleted.‡**
13.	Cited theories and evidence to inform how race is treated methodologically and support by theory of change? Examples include theories about race or associated constructs (e.g., racism) along the causal pathway and treatment in study methods.	**Deleted.§**
C. Measurement Error		
14.	Discussed potential participant sources of measurement error related to racial measurement tool? Examples include racial switching, non-identification with race, self-censor due to social desirability or internalized stigma, influenced by study administrator, and motivations (e.g., study eligibility and compensation).	**Revised.*** Q2) *Potential participant sources of measurement error in race data collection?*

15.	Discussed potential sources of measurement error related to racial measurement tool? Examples include survey options for race (e.g., ability to select all that apply), self-identified qualitative option, inclusion of race/ethnicity as a single question, lack of strong validity evidence, and whether single tool used for all study participants.	**Revised.** Q3) *Potential sources of measurement error due to the race data collection tool(s)*
16.	Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool? Examples include order that question(s) appears overall in survey, single instance of racial data collection, influence of study administrators on response, social stigma, and political climate.	**Deleted.**†
17.	Discussed limitations of assessing measurement error due to an unknown "true value" for race?	**Deleted.**†
	D. Data Reporting	
18.	Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other")	**Deleted.**
19.	Reported rate of racial switching or changes in participant racial classification?	**Deleted.**
	E. Data Analysis	
20.	Justified combining survey options or combinations of race variable in analysis that differ from instrument?	**Revised.*** Q13) *Justification to combine, exclude, or change original race data reporting*
21.	Discussed racial subgroup analysis? Examples might include intersectional discussions of race and gender, gender expression, income, ethnicity, cultural beliefs and practices, strength of racial identity and importance.	**Revised.*** Q8) *Heterogeneity within racial groups*
22.	Discussed testing of statistical assumptions and implications for race? An example includes that the assumption of independence means observations within the same race are unrelated.	**Revised.*** Q15) *Limitations of statistical reasoning due to a race variable*
23.	Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	**Deleted.**‡
24.	Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	**Deleted.**§
25.	Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities? For example, a background section that describes racism as affecting certain groups may mean that a nested model may be more appropriate.	**Deleted.**‡
26.	Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	**Revised.*** Q16) *Interpretability of data results on racial group analysis*
27.	Discussed implications for internal validity due to treatment of race variable analytically? Examples include increased type I and type II error, diminished power, and decreased external validity.	**Revised.*** Q18) *Limitations of external validity due to analytical treatment of race*
	IV. External Validity	
28.	Discussed threats to external validity considering race-related threats to internal validity? For example, measurement error due to lack of construct validity limits generalizability.	**Revised.** Q17) *Limitations of external validity due to the construct or meaning of race used in study (validity)*
29.	Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	**Revised.*** Q19) *Limitations of external validity due to within-group racial heterogeneity*

30.	Discussed other threats to external validity for generalizability claims related to race? Examples include population sameness across geographic areas, claims of temporal validity without accounting for changes in race data collection (e.g., changing Census) and related movements (e.g., Black Lives Matter)	**Deleted.‡**

***** This sentence was revised to improve clarity and readability by reducing vocabulary load, sentence construction, and limiting the scope of the question
† This sentence was deleted because it was redundant with another question
‡ This sentence was deleted because the scope of the question was too broad and imprecise
§ This sentence was deleted because it was too narrow in scope and did not match the level of analysis with other questions. This scope of this question is also subsumed under another question in the Critical Race Framework 2.0.

Table 47: Comparison of the Critical Race Framework 1.0 and 2.0

CR Framework 1.0	CR Framework 2.0
Reliability	
1. Discussed reliability of racial measurement tool?	1. Reliability evidence of survey tool(s) used to collect racial identity
2. Cited prior reliability evidence?	2. Potential participant sources of measurement error in race data collection?
3. Supported reliability of racial measurement tool with multiple measurement?	3. Potential sources of measurement error due to the race data collection tool(s)
	4. Existence of a "true value(s)" for race
Validity	
4. Discussed what construct or meaning race intended to measure?	5. Construct or meaning of race used in study (validity)
5. Discussed relevance of racial admixture to construct or meaning of race used in study?	6. The inclusion of multiracial identity to the construct or meaning of race used in study (validity)
6. Discussed whether a "true value" exists for race?	7. Characteristics intended to differentiate racial groups (discriminant validity)
7. Cited validity evidence for racial measurement tool?	8. Heterogeneity within racial groups
8. Discussed characteristics intended to distinguish racial groups?	
9. Assessed whether racial measurement tool correctly and consistently distinguished races?	
Internal Validity	
10. Assessed threats to internal validity due to race variable? Examples include measurement error, multicollinearity, and racial differences in history, mortality, and group selection.	9. Potential threats to internal validity due to quality of reliability and validity of the race variable alone
11. Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable?	10. Population data estimates for all possible combinations of race based on race data collection tool(s)
12. Discussed limitations of race to establish quality of study design? Examples include data limitations to compare to known population (e.g., lack of racial admixture data), establishing success of random assignment, differences in racial data collection/tools with sample and study populations.	11. Methods to provide participants with study construct or meaning of race during data collection
13. Cited theories and evidence to inform how race is treated methodologically and support by theory of change? Examples include theories about race or associated constructs (e.g., racism) along the causal pathway and treatment in study methods.	12. Data results of all possible combinations of race based on original race data collection tool(s)

14. Discussed potential participant sources of measurement error related to racial measurement tool? Examples include racial switching, non-identification with race, self-censor due to social desirability or internalized stigma, influenced by study administrator, and motivations (e.g., study eligibility and compensation).	13. Justification to combine, exclude, or change original race data reporting
15. Discussed potential sources of measurement error related to racial measurement tool? Examples include survey options for race (e.g., ability to select all that apply), self-identified qualitative option, inclusion of race/ethnicity as a single question, lack of strong validity evidence, and whether single tool used for all study participants.	14. Meeting statistical assumption of independence considering racial grouping
16. Discussed potential sources of measurement error other than sources arising from participants or racial measurement tool? Examples include order that question(s) appears overall in survey, single instance of racial data collection, influence of study administrators on response, social stigma, and political climate.	15. Limitations of statistical reasoning due to a race variable
17. Discussed limitations of assessing measurement error due to an unknown "true value" for race?	16. Interpretability of data results on racial group analysis
18. Reported data frequencies for each stage of research before combining races or excluding certain options from analysis (e.g., non-response, "other")	
19. Reported rate of racial switching or changes in participant racial classification?	
20. Justified combining survey options or combinations of race variable in analysis that differ from instrument?	
21. Discussed racial subgroup analysis? Examples might include intersectional discussions of race and gender, gender expression, income, ethnicity, cultural beliefs and practices, strength of racial identity and importance.	
22. Discussed testing of statistical assumptions and implications for race? An example includes that the assumption of independence means observations within the same race are unrelated.	
23. Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?	
24. Made adjustments in analysis to account for error in race variable to prevent Type I and Type II errors?	
25. Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities? For example, a background section that describes racism as affecting certain groups may mean that a nested model may be more appropriate.	
26. Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?	
27. Discussed implications for internal validity due to treatment of race variable analytically? Examples	

include increased type I and type II error, diminished power, and decreased external validity.	
External Validity	
28. Discussed threats to external validity considering race-related threats to internal validity? For example, measurement error due to lack of construct validity limits generalizability.	Limitations of external validity due to the construct or meaning of race used in study (validity)
29. Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis?	Limitations of external validity due to analytical treatment of race
30. Discussed other threats to external validity for generalizability claims related to race? Examples include population sameness across geographic areas, claims of temporal validity without accounting for changes in race data collection (e.g., changing Census) and related movements (e.g., Black Lives Matter)	Limitations of external validity due to within-group racial heterogeneity
	Limitations of external validity due to social and political changeability of race

2.A) Scale. There is no scientific consensus on the appropriate scale for critical analysis or bias tools for research reviews. Some scales include yes-no options while others include qualitative assessment on a low-high scale (Eldridge et al., 2016; Leary et al., 2011; Munn et al., 2014). Initially, the scale for the CR Framework 1.0 was primarily designed to ascertain whether the research article satisfied the condition of the question prompt, such as "discussed…" ("yes") or did not discuss ("no"). This scale for Critical Race Framework 1.0 was derived from the literature – the Joanna Briggs Institute Prevalence Critical Appraisal Tool (Munn et al., 2014). The CR Framework 1.0 was similar to this tool because it was intended for assessment of research quality and usefulness (Munn et al., 2014). The CR Framework 1.0 had an ostensible major advantage over this tool. Rather than ask study participants to respond to highly subjective prompts or give opinions, the CR Framework intended to ask verifiable questions, meaning that there was a purported correct answer. The training specifically instructed study participants to disregard the quality of the discussion, the quality of evidence, or the number of citations. They were asked to indicate ("yes", "no", "unclear", or "not applicable") whether such as discussion or citation was indicated (e.g., "Discussed reliability of racial measurement tool?"). The PI had

intended that this analytical method would reduce survey fatigue, which is associated with suboptimal survey quality and premature study termination (Lavrakas, 2008). The Critical Race Framework 1.0 was designed to be factual rather than scorer-dependent or opinion-based.

The average distributions across questions for the CR Framework were "yes" (42%) and "no" (45%)(Table 24). The PI's scores of the Phase I article were as follows: "yes" (n=10, 33%), "no" (n=17, 57%), "not applicable" (n=3, 10%). However, the disagreement among senior faculty with decades of research experience made a major impression on a decision whether to continue with the original scale – 97% of "no" responses versus 100% of "yes" responses. A doctoral student remarked, "Consider a scale (e.g., 1-5, instead of dichotomous measures)?" (Table 26). This was the only feedback on the scale itself. The remaining respondents had varying responses based on their most common selection (n=14-18, 47%-60%).

The PI concluded that applying the CR Framework was inherently a qualitative exercise. As much as this study had intended to determine whether the research article used in Phase I "discussed," "cited," "reported," "justified," "interpreted," "applied," or "assessed" topics, the CR Framework 1.0 relied on individual expertise and subjective judgment. In conjunction with the initial scales that are discussed above, the PI reviewed the Cochrane Collaboration's tool for assessing risk of bias in randomized trials (Eldridge et al., 2016). The CR Framework 2.0 for Phase II was modified to include options for "high," "moderate," "low," and "no discussion." This scale adopted the high-low assessment used in several critical analysis tools (Eldridge et al., 2016; Leary et al., 2011).

2b) Question Wording. Sentence construction varied in the Critical Race Framework 1.0. Whereas some questions were succinct and straightforward (e.g., "discussed reliability of racial measurement tool"), many questions contained complex sentence construction. An aid

proceeded the actual question (e.g., “discussed limitations of race to establish quality of study design? Examples include (*underlined for emphasis*) data limitations to compare to known population..”) for ten out of thirty questions.

All questions in the Critical Race Framework 1.0 began with an action verb. In conjunction with a yes/no scale, this primed users to think dichotomously. We concluded that this structure may create a substantial source of measurement error due to the potential for a false set of choices. Yes/no questions can force respondents to imprecisely or inaccurately select responses due to limited options (Choi & Pak, 2005). The Critical Race Framework 2.0 avoided the use of action verbs and opted for topic areas (e.g., “characteristics intended to differentiate racial groups”). This change also addressed feedback from a Phase I participant, “Be careful and intentional about discuss vs assess and communicate to rater what the difference is. In some instances, it seemed to me that authors may have discussed, though not assessed, a certain item.” (Table 26). The PI has not intended to draw such a distinction in word choice. The original framework used nine different verbs that may have introduced ambiguity for users.

Questions containing examples (e.g., “Examples include order that question(s) appears overall in survey, single instance of racial data collection”), may have been confusing to framework users and reduce readability, defined here as the vocabulary load and sentence construction. Changes to the framework were intended to avoid jargon and employ simple sentence structures consistent with readability standards (DuBay, 2004).

There was a single comment on the sentence construction, “Be careful of double-barreled items.” Phase I respondents indicated that question clarity needed improving (Table 26). The PI reviewed each question for clarity, sentence structure, and redundancy. Questions that separately asked about discussion and citations on the same topics were consolidated, as with question 1

("Discussed reliability of racial measurement tool?") and question 2 ("Cited prior reliability evidence?"). The new question became, "Reliability evidence of tool(s) used to collect racial identity". Some questions that were targeting the same scientific construct were also consolidated, as with these questions on discriminant validity (i.e., Q8. "Discussed characteristics intended to distinguish racial groups?" and Q9. "Assessed whether racial measurement tool correctly and consistently distinguished races?").

2c) Contextual Cues – We streamlined the CR Framework so that definitions appeared in a separate column – the User Aid section. This served several purposes. First, it provided a contextual cue to clarify the word or phase being used in the framework. This helped to minimize survey or respondent bias, so that there was a shared meaning. It also benefited the actual CR Framework question, which was now free of additional text.

2d) Length – The length of the Critical Race Framework was a major concern based on data results from Phase I. The average time to read and understand the Critical Race Framework 1.0 was 21-40 minutes. Across four discrete activities for Phase I (reading CR Framework, web-based training and Part 1 survey, review assigned article, and Part 2 survey), the average Phase I participants spent was considerable (>1 hour), as discussed more fully in the results section.

2e) Revised Questions from the Critical Race Framework 1.0. The following discussion provides the rationality for changes from the Critical Race Framework 1.0 to 2.0. It is lengthy, but no exhaustive of all changes. However, iteration of the Framework derived from systematic approaches consistent with our conceptual model (Figure 2).

1) Discussed reliability of racial measurement tool? (Q1) This question was revised to, "Reliability evidence of survey tool(s) used to collect racial identity" (Q1). The use of the word measurement was inappropriate in this context of this study since measurement connotates

quantification of attributes that is usually of an object or event. Rather, “survey tool” is consistent with established language for qualitative and quantitative research. It appears 99,200 times in the literature based upon a word phrase search (“survey tool”) conducted in Google Scholar on July 13, 2023. It was also pluralized because studies may employ multiple tools to collect racial identity or combined different data sets for data analysis. Only 38% (n=3) of Phase I participants scored the original question as “highly relevant” to reliability, which is attributed to the use of measurement rather than survey tool. One participant, a faculty member, indicated difficulty understanding, is attributed to the lack of question clarity. This participant was “very familiar” with the construct of reliability. The study definition of reliability as the ability of the race data collection tool to provide consistent responses for participants’ racial identity or identities appears in a middle column of the Critical Race Framework 2.0 (Heale & Twycross, 2015).

2) Discussed what construct or meaning race intended to measure? (Q4) This question was revised to, “Construct or meaning of race used in study (validity)” (Q5). Since this is the definition of validity, no further changes were needed other than to improving reading efficiency with “used in study” rather than “intended to measure.” As discussed earlier, the word *measure* connotes quantification that is inappropriate because race is a qualification. This may explain why only 38% (n=3) of respondents indicated that the original question in the framework was “highly relevant” to validity. Despite the drawbacks of “measure,” the question otherwise matched the definition of validity. The low percentage for “highly relevant” may also reflect varying levels of familiarity and expertise among participants and provide evidence that the Critical Race Framework is limited to experienced quantitative users.

There was no general agreement among teaching faculty (n=4) as to the relevance to validity ("highly relevant" = 2, "fairly relevant" = 1, "somewhat relevant" = 1)(not shown). This disagreement may reflect a research position for some that race does not require a related construct. Race without qualification is a widely accepted research practice [citation]. This study argues that without a related construct for race, scientific quality is weakened.

3) Discussed relevance of racial admixture to construct or meaning of race used in study (Q5). This question was revised to "The inclusion of multiracial identity to the construct or meaning of race used in study (validity)" (Q6). Race is not mutually exclusive, which problematizes discriminant validity for race construction. Racial admixture was removed because this phrase is less common than multiracial identity.

4) Discussed whether a "true value" exists for race? (Q6) This question was revised to "existence of a "true value(s)" for race". There were minor changes associated with the new question. "Discussed" was abandoned for all questions. "True value" was pluralized to account for the possibility of having multiple "true" races. This question garnered two responses for "difficulty understanding" and two for "should be reworded". The responses on difficulty are attributed to a lack of familiarity with the scientific reasoning on "true value" for some participants. The study recruited researchers with diverse training (e.g., PhD, DrPH, EdD, MPH) and research backgrounds (e.g., teaching faculty, non-teaching faculty, doctoral student). True value is an inviolable assumption for assessing measurement error and statistical interpretation (citation). Measurement error can be estimated only if a true value exists (citation). Since statistical testing accounts for random error due to chance, testing results should be interpreted cautiously. Thus, the question was retained since the notion of true value is essential to evaluating bias and data results.

This question was also tied for the lowest scored question on relevance to related core construct. In the Critical Race Framework 1.0, On average, participants scored it a 2.8. Several theories could explain this study result. First, the relevance of true value may be more appropriate for another related core construct – reliability, internal validity, or external validity. Second, study participants may believe that the notion of true value for race is not germane to article analysis. Use of race without critical reflection on its threats to research quality is common in the public health literature [citation]. Third, some participants may be less familiar with quantitative reasoning and, thus, unclear on the relevance of this question. In fact, the answer to whether there exists a "true" race(s) for any given study is important for each of the four core concepts. Its lack of mention in a study weakens an ability to evaluate consistency (reliability), define a construct or race (validity), accurately interpret data findings (internal validity), and generalize to a population (external validity). The PI moved this question from validity to reliability because it is fatal to reliability and, thus, fatal to validity based on the theoretical framework of this study (Graph). If retained under validity as in the Critical Framework 1.0, users may overlook the fatal threats to reliability by assuming that reliability can be established without a true value. Reliability is subsumed under validity. By implication, validity cannot be established if no true value exists for race. The Critical Race Framework underscored this point to users by adding, "Measurement error cannot be assessed without a true value" to the key concepts section.

5) Discussed characteristics intended to distinguish racial groups? (Q8) This question was revised to, "Characteristics intended to differentiate racial groups (discriminant validity)" (Q7). Minor changes to the original question included removal of "discussed" and substitution of distinguish for differentiation. The first general definition in the Merriam-Webster dictionary

lists differentiate as “to mark or show a difference in: constitute a contrasting element that distinguishes” and distinguish as “to mark as separate or different”. Both definitions refer to the other, meaning that they are synonyms.

`However, additional definitions of distinguish were considered. The PI conducted a Google search on July 14, 2023, for “definition distinguish”. The first result, which Google derived from Oxford Languages dictionary, defined distinguish as, “recognize or treat (someone or something) as different [citation]. Treatment or mistreatment was beyond the scope of what this study question had intended, which was to emphasize difference or discriminant validity. This study concluded that difference is a more suitable term for the purposes of this study.

6) Assessed threats to internal validity due to race variable? Examples include measurement error, multicollinearity, and racial differences in history, mortality, and group selection (Q10). This question was revised to, “Potential threats to internal validity due to quality of reliability and validity of the race variable alone” (Q9). As previously discussed, all explanations of key concepts and examples were moved to the user aid column. The scope of this question was narrowed to assess the user’s overall determination of the study’s quality of discussion of internal validity due to the quality of reliability and validity. The qualifier (“of the race variable alone”) was a cue for users since this question may be susceptible to misinterpretation or broad application beyond the reliability and validity of race, meaning users may extrapolate the question to a general judgment of the study’s quality of discussion on reliability and validity. However, the scope of this question remains sufficiently broad consistent with the intent of primary questions used in this study.

7) Discussed limitations of causal inferences between treatment and observed outcomes due to threats to internal validity from race variable? (Q11). This question was revised to,

"Limitations of statistical reasoning due to race variable" (Q15). Consistent with changes to the Critical Race Framework, a topic was substituted for a question-based prompt. Sixty-seven percent (n=4) of Phase I respondents indicated that the selected article discussed limitations of causal inferences between treatment and outcomes. Two respondents provided "No" (17%) and "Unclear" (17%) responses. No participant determined that the question was not applicable. The PI's decision that this question was not applicable ("not applicable") was based on the study design – no treatment was used in the study by Arul & Mesfin (Arul & Mesfin, 2017). Treatment is associated with experimental designs, but their study design was cross-sectional. The differences among Phase I participants and between PI and participants are attributed to subjective interpretation. "Treatment" may have been more liberally interpreted to mean the independent variable. To account for different study designs and improve the applicability of the Critical Race Framework, statistical reasoning was deemed more appropriate to align with the purpose of the question – to bound statistical interpretation to the potential limitations introduced by a race variable.

8) Discussed potential participant sources of measurement error related to racial measurement tool? Examples include racial switching, non-identification with race, self-censor due to social desirability or internalized stigma, influenced by study administrator, and motivations (e.g., study eligibility and compensation) (Q14). This question was revised to, "Potential participant sources of measurement error in race data collection" (Q2). Supplemental information to guide readers was moved to the user aid. Although "measurement" garnered critical appraisal earlier in the discussion, "measurement error" is appropriate here because it is a common scientific term related to research analysis. The deletion of "racial measurement tool" avoided an implied quantification of race. The revised question was moved under reliability

since it concerned “error in race data collection.” The original question received an above average rating (3.5) on relevance (to validity). That question was imprecise on the activity or context where measurement error might arise. Consistent with the narrower scope of secondary questions, the new question was limited to the data collection process (e.g., completing survey).

9) Discussed potential sources of measurement error related to racial measurement tool? Examples include survey options for race (e.g., ability to select all that apply), self-identified qualitative option, inclusion of race/ethnicity as a single question, lack of strong validity evidence, and whether single tool used for all study participants (Q15). This question was revised to “potential source(s) of measurement error due to the race data collection tool(s)” (Q3). The justification for this change was similar to 8) above for removal of “measurement tool” and placement of examples in user aid. This question was retained due to its importance to the essential problem statement that this study premises.

10) Justified combining survey options or combinations of race variable in analysis that differ from instrument? (Q20). This question was revised to, “Justification to combine, exclude, or change original race data reporting” (Q13). This question was tied with several questions for the highest relevance score (3.5). This score reflects the relevance of that question to the related construct. A single user indicated “difficulty understanding” this question. This participant was a teaching faculty member with considerable research who was unclear on the instructions. This person left feedback in the open comments at the end of the survey, “But I then found the survey afterwards to be really confusing, as I was not sure if the questions were in relation to the CR framework AND the paper or just the (C)R” (Table 26). The study participant found that the survey had a dual meaning. This feedback was unique to this participant. No other participant raised this concern.

The instructions for this section of the Part II survey read as, “Now, we would like feedback on the relevance of each question of the CR Framework to the concept.” In the introduction for the relevance section on internal validity, it reiterated, “Please assess the relevance of each question to the concept of Internal Validity considering the race variable alone.” An explanation for the participant’s confusion was not readily apparent other than not carefully reading the instructions. The user’s feedback could be attributed to survey fatigue, but that is not likely. Two participants completed Phase I in less than one day. This user had five days between Part I and II. A more likely explanation is that the timing of the study coincided with the end of the academic year. This participant’s time completion for Phase I was well below the average (8 mins versus 26 mins). This person’s comment may indicate rushed completion of survey arising from competing priorities at the end of the semester.

11) Discussed racial subgroup analysis? Examples might include intersectional discussions of race and gender, gender expression, income, ethnicity, cultural beliefs and practices, strength of racial identity and importance (Q21). This question was revised to, “Heterogeneity within racial groups” (Q8). The original question was wordy and lengthy. Descriptions were revised and moved to the user aid. The question was also moved from the internal validity to the validity section due to its below-average (3.0) relevance score. The absence of a discussion on within-group diversity is applicable to reliability (e.g., lack of practical significance, measurement error), validity (e.g., poor or biased construct, measurement error), internal validity (e.g., alternative explanations), and external validity (e.g., limited generalizable without consideration of racial heterogeneity). The decision to place the revised question under validity is to draw attention to the biasing and limitations of monolithic racial grouping in race construction and to the threat to validity. As constructed, “within racial groups”

could be constructed as supporting race construction of which this study is highly critical. However, the effect on future respondents is expected to be minimal, if not absent, because the Critical Race Framework 2.0 underwent bias screening during development, as described above. In addition, the framework is intended as a critical analysis tool for research articles that use racial classification. The Critical Race Framework 2.0 does not introduce a bias in reference to racial groups given that race is ubiquitous in public health research and common parlance. It is unlikely that user extrapolated this question to a value-judgment about the soundness of racial construction given the otherwise critical reflection question in the framework.

12) Discussed testing of statistical assumptions and implications for race? An example includes that the assumption of independence means observations within the same race are unrelated (Q22). The question was revised to, "Limitations of statistical reasoning due to a race variable" (Q15).

13) Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results? (Q26). This question was revised to "Interpretability of data results on racial group analysis" (Q16). The original question was wordy and double-barreled (e.g., accurately interpreted v-value and account for alternate explanations). The *p-value* is too limiting for the purposes of the Critical Race Framework and, thus, was eliminated in the revised question. *P-value* is one of many measures, such as confidence intervals and model fit, that can be used to help interpret statistical significance. Though not an exhaustive list, other measures include confidence interval, effect size, correlation coefficient, and chi-square value. In addition, *P-value* concerns the probability of random error or the error due to chance and does not account for errors or bias in conceptualization and testing or limitations in practical significance.

14) Discussed implications for internal validity due to treatment of race variable analytically? Examples include increased type I and type II error, diminished power, and decreased external validity. (Q27) This question was revised to "limitations of external validity due to analytical treatment of race" (Q18). It is a secondary question for external validity. Its relevance score was average (3.3). One study participant – an experienced teaching faculty – felt that this question was redundant. The PI agreed with this feedback. This question was similar to question 25 ("Applied appropriate data analysis based upon background or discussion on explanations for racial health disparities?"), question 26 (" Accurately interpreted p-value to account for alternate explanations associated with race (e.g., effects of racism) that may explain study results?"), and question 23 (" Discussed analyses involving systematic error (e.g., measurement error) and random error (e.g., error associated with chance) related to the race variable?". The basis for each of these questions concerned drawbacks for establishing internal validity due to the use of race in analysis. Questions 23 and 25 were eliminated in the Critical Race Framework 2.0 to reduce redundancy.

The original question was under internal validity and had intended to draw attention to the effects of race analytical treatment on Types 1 and II errors, power, and external validity. This question was triple-barreled, as read in full. The inclusion of external validity in this question on internal validity appears to be an error in survey development. Although they raise important considerations for assessing any research article, a goal and the scope of the Critical Race Framework were to reduce redundancy, enhance question clarity consistent with its purpose, and balance the scale. The PI chose external validity from among the options in the Critical Race Framework 1.0.

15) Discussed threats to external validity considering race-related threats to internal validity? For example, measurement error due to lack of construct validity limits generalizability (Q28). This question was revised to, "Limitations of external validity due to the construct or meaning of race used in study (validity) (Q17). This question narrowed the focus to the threats to internal validity arising from the study construction of race. As a primary question, it was intended to be broader in scope than secondary questions. However, the initial question was too broad and imprecise, which impacts the quality and interpretability of survey data findings.

16) *Assessed external validity considering racial heterogeneity, intersectionality, and subgroup analysis* (Q29). This question was revised to, "Limitations of external validity due to within-group racial heterogeneity" (Q19). The question that appeared in the Critical Race Framework 1.0 was triple-barreled – to assess racial heterogeneity, intersectionality, and subgroup analysis. Only racial heterogeneity was retained for the revision.

3. Refinement of the Critical Race Framework Training

Feedback on the web-based training was not uniform. Some provided highly positive feedback ("I think this training should be replicated. It was clear, concise, informative, and advanced, at the same time" and "The videos were helpful and clearly explained each section of the CR Framework.") while others had constructive feedback ("The training was too long… The transcript was not verbatim." and "Consider a more interactive training."). We identified several goals for this stage: retaining the adult learning theoretical underpinnings, ensuring a verbatim manuscript, shortening the training video, creating a more professional product, and removing module-based training.

Consistent with applied adult learning theory in Phase I, we identified four relevant areas: spotting potential issues in research, refining their own research techniques ("personal payoff"),

and appealing to personal motivation and lifelong learning. It also integrated principles that adult learners should be involved in the content development and evaluation by encouraging study participants to sharing their feedback and reflection throughout study participation, which supports ongoing improvement of the CR Framework and related training. The introduction also addressed Bryan and colleagues' guidance for adult learning in public health practice by involving learners as stakeholders through ongoing improvement and define a key aspect of public health practice to assess the quality of research quality (Bryan et al., 2009). Participants were asked to provide extensive quantitative and qualitative feedback on the Critical Race Framework training for improvement consistent with adult learning theory.

We overhauled the training video to involve artificial intelligence (AI) models to deliver instruction. The PI provided a slightly shorter transcript than appeared in Phase I for ten AI models using the Heygen application within Canva – an Internet digital design website (*HeyGen - AI Video Generator*, n.d.; *Home*, n.d.).

---

5. Research Question 3

RQ3.What are demographic and research associations with perceptions of the Critical Race Framework? (Phase II-III)

Finding: *Due to a small sample size, we cannot make any generalizable statements about factors associated with perceptions of the CR Framework, only that our data did not generally show statistically significant associations with demographic/research factors and measures of fit.*

The sample size for Phase II was not large ($n \approx 22$). However, our tests were robust to small sample sizes. We discuss our findings from the Mann–Whitney *U* tests and correlation analyses. We confidently accepted our study results that demographic or research factors were

not generally associated with perceived *measures of fit* – acceptability, appropriateness, feasibility, and satisfaction. There is no evidence that differences due to experience and demographics were introduced in our study.

The Mann–Whitney *U* test tests the null hypothesis that two groups are drawn from the same population – "homogeneous and have the same distribution" (Nachar, 2008). Its strengths as a nonparametric test are considerable, "This test has the great advantage of possibly being used for small samples of subjects (five to 20 participants). It can also be used when the measured variables are of ordinal type and were recorded with an arbitrary and not a very precise scale" (Nachar, 2008). The Mann-Whitney *U* test does not assume distributional properties (e.g., normality). In the presence of "one or two extreme values," it loses little statistical power and can produce accurate statistical results (Nachar, 2008). The Mann-Whitney *U* test may be less reliable "whenever one's samples are drawn from two populations with a same average but with different variances" (Nachar, 2008). Our dataset was small and robust for the Mann-Whitney *U* test. We concluded that our data results strongly suggested no differences in employer type, primary role, dichotomized public health specialty, and type of terminal degree in *measures of fit* between partial and complete response groups.

Except for two tests, we did not find statistically significant Kendall coefficients. These coefficients are interpreted as the "probability that any pair of observations will have the same ordering on both variables rescaled to range from -1.0 to 1.0" (Arndt et al., 1999). Kendall's tau can accommodate small sample sizes with 30 or fewer and superiority over parametric tests across most sample sizes. In tests of eight samples, Kendall's correlations can perform well (Arndt et al., 1999). We concluded that our results can be reliably interpreted. However, we ascribed these findings in perceived appropriateness to two survey respondents. Unless otherwise

noted, the number of years of holding a public health doctoral degree, number of years in public health research and public health teaching, familiarity with research methods, self-assessed research skills, and the level of agreement with the study premise are not associated with perceptions of acceptability, appropriateness, feasibility, and satisfaction with the CR Framework.

Based on the study results, our RQ3 findings are not explained by potential differences among our experts. However, there are important considerations. We previously concluded that Phases I and II data are MNAR and that partial and completes cases did not differ in four areas (i.e., type of doctoral degree, type of employer, employee type, and public health specialty) (Table 35). A determination of MNAR meant that there were unobservable factors associated with attrition. As such, we cannot generalize our results to a general population of public health experts. Future research should support improvements in study design, participation incentives, and recruitment strategies to establish external validity. Populations less likely to enroll or fully participate in this study may differ from our sample population.

Our results provided little evidence to guide enhanced recruitment and retention strategies. We initially speculated that the Critical Race Framework training could explain attrition, but our Qualtrics user data analytics showed that most who dropped out in Phases I and II never saw the Critical Race Framework training page or tool. A monetary incentive, as opposed to the lottery that we used in Phase II, might support higher retention. We did not follow-up with non-respondents – a weakness in this study. Any reasons to explain low enrollment and high attrition are highly speculative but can be explored in future research. This study design may be more feasible with an existing critical appraisal group or organization.

6. Research Question 4

What is the a) reliability and b) validity evidence for the Critical Race Framework? (Phase II-III).

Finding: a) Reliability – *Interrater agreement was moderate to high among rater pairs, which does not fully establish interrater reliability, but shows promise. Due to lack of confidence in significance testing, we found that interrater reliability results were inconclusive.*

In our conceptual framework, we defined eight realms of inquiry derived from reliability theory (Table 3). An evaluation of reliability of any given study requires qualitative and quantitative judgment. Throughout each section in Chapter 5, we integrated applied theory in our discussion. Here, we interpreted quantitative data: non-differentiation and interrater reliability analysis.

The ICCs for our adopted scales yielded excellent reliability, consistent with the literature: acceptability (Cronbach's $\alpha = 0.9$), feasibility (Cronbach's $\alpha = 0.87$), and appropriateness (Cronbach's $\alpha = 0.91$)(Table 37) (Proctor et al., 2011). Although these scales were previously validated, our study findings support the underlying construct validity from which we based our findings. Our analyses were robust to small sample sizes.

We conducted non-differentiation analysis in Phase I and II. We did not conduct non-differentiation analysis in Phase II because we anticipated low ratings, consistent with the literature, that would not enable interrater reliability testing to run. In Phase I, we did not find strong evidence of scale non-differentiation: acceptability (0.5), feasibility (0.33), appropriateness (0.5), relevance (0), and framework (0.17). We found that a single user likely exhibited straight-lining during the article critique phase – the same response ("yes"). However, we found that Phase I data could not be reliably interpreted due to low enrollment, high attrition,

small sample size, and major needed improvements in study design, instrumentation, and training. Our finding of non-differentiation is not meaningful in the broader determination of poor reliability in Phase I.

In Phase II, we did not find strong evidence of scale-level non-differentiation with values ranging from 0.10 to 0.67. Two individuals or 10% of the study sample likely exhibited straight-lining for questions on relevance. Each selected the same responses for all twenty questions. Their responses did not constitute a sufficient threat to reliability of our study results. Although our non-differentiation finding was similar for Phase I, Phase II had major strengths due to higher enrollment (i.e., 266% increase), more stringent eligibility screening (e.g., only individuals with terminal degrees), and major improvements in study design, instrumentation, and training.

Patterns in pairwise rater agreement was more consistent in high/moderate versus low/no agreement as opposed to absolute agreement. Raters 1 and 2 agreed on 80% or more articles per item 14 out of 20 times using the dichotomized scale (Table 43). Raters 1 and 3 agreed for 80% or more articles per item 18 out of 20 times. The significance testing for our weighted Kappa coefficient cannot be reliably interpreted due to sample size. Although weighted kappa coefficients have an advantage in correcting for chance over crude agreement, we did not meet minimum size requirements. “Weighted kappa can be validly applied when the number of subjects being rated, n, is greater than $2c^2$ where c is the number of categories” (Soeken & Prescott, 1986). Our “subjects” were articles, which numbered 10 in each analysis. The number of categories was 4. We would have needed at least 32 articles for valid results.

As we previously described, ten article reviews have been used in the literature (Kmet et al., 2004). We were limited to ten because of raters’ time constraints and study delimitations.

Future research should seek to study and reduce barriers to ensure that minimum sample sizes can be met for interrater reliability testing. We also encourage intra-rater reliability testing, which would involve raters evaluating articles at different time periods. Two months between data collection would be optimal to assess intra-rater reliability.

Finding: b) Validity – *Overall data results showed excellent content validity. Our results were inconclusive on internal validity and external validity. Evidence of construct validity for reliability and validity items in the CR Framework was poor to fair yet promising. We consider our results as exploratory, but a contribution to the literature nonetheless because it provides preliminary data against which future research can confirm or reject.*

In our conceptual framework, we defined eight realms of inquiry derived from validity theory (Table 4). We carefully considered each realm of inquiry throughout all stages of our study. Implications of reliability and validity in data analysis and interpretation are discussed in each section in Chapter 5. In this section, we discuss results of our validity (Phase II) and reliability (Phase III) testing.

Validity Evidence

1. Content Validity

Overall data results showed excellent content validity. Item- and scale-level CVIs were excellent. The single exception was "existence of a 'true value' for race" (I-CVI=0.70). Our Kappa values accounted for chance agreement and yielded no significant change from I-CVIs. In their highly work on content validity and CVI, Polit and colleagues concluded that, "items with an I-CVI of 0.78 or higher for three or more experts could be considered evidence of good content validity" (Polit et al., 2007). Our study attracted 22 experts. 17 out of 20 I-CVIs exceeded 0.78. Polit and colleagues elaborated, "(O)ur recommendation falls somewhere in

between the conservatism of a minimum S-CVI of .80 for the universal agreement approach and the liberalism of a minimum S-CVI of .80 for the averaging approach" (Polit et al., 2007). Our S-CVIs all met or exceeded this threshold. A content validity study of the mixed methods appraisal tool (MMAT) had an acceptable relevance agreement index for 6 of 21 items (29%) (Hong et al., 2019). Our study is notable because, unlike many critical appraisal studies like the MMAT that relied on Delphi techniques in tool development, a single PI (CW) directed the construction and revision of the CR Framework. Considering our study findings on CVI, this is a major achievement.

As discussed in Chapter 2, the notion of a "true value" is critical for assessing measurement error, which is the difference between a measured observation and its true value. If race is not assumed to have a true value, then measurement error of a race data collection tool is indeterminant. This item in the CR Framework may have posed difficulty for some participants because race is widely assumed to be stable (Agadjanian & Lacy, 2021). Further, as Root explained, "The statistician does not consider a third option: that a member's true race is not the race she reports herself to be on any survey" (Root, 2008). In other words, norms on race conceptualization may explain the single instance with a low I-CVI. Future training and research would benefit from creating training materials to discuss and affect research norms that may give rise to relevance scoring.

Further research is needed in content validity. A limitation of this study is that we did not use criterion validity, regarded as a "gold standard" (Cordier et al., 2023). Given the four-pronged framework of this study, each representing singular rich fields of study, we did not benefit from an existing set of criteria. A larger and more representative sample size would enhance content validity. Many measures in our study on demographics and research experience,

however, may have limited population estimates for determining representativeness. We viewed our sample size as robust for content validity analysis. Five experts are considered the minimum sample size for content validity (Larsson et al., 2015; Lynn, 1986).

2. Construct Validity

Our results were inconclusive on internal validity and external validity. Evidence of construct validity for reliability and validity items in the CR Framework was poor to fair yet promising. We consider our results as exploratory, but a contribution to the literature nonetheless because it provides preliminary data against which future research can confirm or reject. There is no general agreement on the threshold for an adequate sample size in use of EFA, though no shortage of recommendations (Beavers et al., 2019; Mundfrom et al., 2005; White, 2022). As a general heuristic, an absolute minimum of 10 observations per variable is required. The participant-to-variable ratio may be more appropriate than the number of variables to determine sample size (Mundfrom et al., 2005). We relied on Beavers and colleagues' guidelines since there is no index in EFA for determining a minimum sample size (Table 11) (Beavers et al., 2019). EFA interpretation relies on qualitative and quantitative judgment. Thus, the stepwise approach from Beavers and colleagues is instructive.

We treated our EFA 3 (internal validity) results with the highest level of scrutiny because the participant-to-variable ratio (3:1) was the lowest. Even assuming excellent criterion (0.98), a fixed variables-to-factors ratio (*p/f*) of 3, a high communality, the minimum sample size (n=32) under guidance from Mundfrom and colleagues cannot be met with our small sample size (n=22)(Mundfrom et al., 2005). We judged that EFA 3 did not have an adequate sample size. The data were not reliable. In addition, we designed the internal validity section to be twice the

length of other sections because it contained more subdomains. Our finding on construct validity for EFA 3 is inconclusive. No further analysis was considered.

Consistent with Beavers and colleagues' guidance, we conducted tests of factorability (Table 39). They devised this approach because there is no general agreement on determining a minimum sample size in EFA. No single factorability index determines whether there was an adequate size, however. We found three determinants that were statistically different from zero: EFA 1 (0.15, P-value = <0.001), EFA 2 (0.06, P-value = <0.001), and EFA 4 (0.10, P-value = <0.001). Each EFA has some linear combinations that make EFA possible. In addition to linear combination, common variance is needed for EFA. The qualitative judgment for our KMOs were middling (EFA 1, 2) and mediocre (EFA 4), respectively 0.7, 0.73, and 0.6 (Kaiser, 1974). At this point, we eliminated EFA 4 from subsequent analysis. A KMO of 0.60 was too low in our estimation to reliably interpret EFA statistics, particularly given our small sample size. It bordered on "miserable."

In EFA 1, the total variance explained for the first factor was 58%. All sums of squared loadings for EFA 1 were acceptable. Since no item loading was less than 4.0, we determined that none were unrelated to the others. Item 3 ("potential sources of measurement error due to the race data collection tool(s)") had the highest loadings (0.99). Our communality pattern was described as "wide," both with and without the lowest loadings. We concluded that EFA 1 statistics provided poor to fair evidence of construct validity for reliability.

In EFA 2, the total variance explained for the first factor was 69%. All sums of squared loadings for EFA 2 were acceptable but cautiously interpreted: 0.92 (*Item 8*), 0.86 (*Item 5*), 0.84 (*Item 6*), and 0.68 (*Item 7*). Our communality pattern was high, which suggests strong data in our EFA. Of all four EFAs, we judged the validity items and section to perform the best – quality

factorability, highest explained variance, and high sums of squared factor loadings. These findings provided poor to fair evidence of construct validity for validity. Our sample size may not be adequate.

As discussed earlier, the Critical Race Framework study would benefit from curricular changes that describe and affect research norms more fully. The brief training on the CR Framework is likely inadequate. Some qualitative data results support this, as discussed. In addition, a consensus approach has been shown effective in increasing agreement (Pace et al., 2012). This option is worth exploring in subsequent research.

Phase II did not involve article critique. Had Phase II resembled Phase I, it may have more favorably influenced perspectives on relevance. However, article critique and survey research in the same study stage may result in low enrollment and high attrition.

EFA results also suggested that our construction of "primary questions" may warrant re-consideration. As we discussed in our findings for RQ2, we had intended "primary questions" to be broader in scope and align more closely to the definition of each construct in the framework. We expected them to have the highest sums of squared loadings, indicating the highest proportion of variation that is explained by the first factor. The primary question did not have the highest sums of squared loading in any EFA. However, we recognize the many limitations in our analyses. Future research should seek to verify our results on the proportion of variance. A larger sample size would be needed.

7. Research Question 5

What is the quality of 1) highly cited health disparities studies and 2) behavioral health studies based on the Critical Race Framework?

Finding: *Twenty studies drawn from the health disparities and behavioral health literature showed low quality or no discussion when the Critical Race Framework was used.*

The articles did not perform well against the Critical Race Framework. Seventy-five percent (n=15) did not have more than 25% of CR Framework items to have "high" or "moderate" quality discussion. These findings are consistent with the literature on the underreporting of racial data conceptualization and justification (Martinez et al., 2022). Since the reliability and validity of the Critical Race Framework remains under development, we cannot be certain of the "true value" of each of these articles' item and overall ratings. Unknown factors may be associated with ratings that should be part of future research. For example, behavioral health articles had lower overall ratings, which may be a function of the two raters. They were recent PhD students in the same academic department in a school of public health. However, we determined the articles from a convenience sampling, as opposed to relying on a bibliographic analysis of highly cited health disparities articles. All raters were from the same institution, which may present a bias. Future research would benefit from more robust reliability and validity analyses and additional raters to determine the quality of ratings using the CR Framework.

Bibliography

Adkins-Jackson, P. B., Vázquez, E., Henry-Ala, F. K., Ison, J. M., Cheney, A., Akingbulu, J., Starks, C., Slay, L., Dorsey, A., & Marmolejo, C. (2023). The role of anti-racist community-partnered praxis in implementing restorative circles within marginalized communities in southern California during the COVID-19 pandemic. *Health Promotion Practice*, *24*(2), 232–243.

Agadjanian, A. (2022). How Many Americans Change Their Racial Identification over Time? *Socius*, *8*, 23780231221098547. https://doi.org/10.1177/23780231221098547

Agadjanian, A., & Lacy, D. (2021). Changing votes, changing identities? Racial fluidity and vote switching in the 2012–2016 US Presidential Elections. *Public Opinion Quarterly*, *85*(3), 737–752.

Ainley, M. (2006). Connecting with learning: Motivation, affect and cognition in interest processes. *Educational Psychology Review*, *18*, 391–405.

Al-Jundi, A., & Sakka, S. (2017). Critical appraisal of clinical research. *Journal of Clinical and Diagnostic Research: JCDR*, *11*(5), JE01.

Allen, K., Zoellner, J., Motley, M., & Estabrooks, P. A. (2011). Understanding the internal and external validity of health literacy interventions: A systematic literature review using the RE-AIM framework. *Journal of Health Communication*, *16*(sup3), 55–72.

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221.

American Society of Human Genetics. (2018). ASHG denounces attempts to link genetics and racial supremacy. *American Journal of Human Genetics*, *103*(5), 636.

Andersen, R. E., Crespo, C. J., Bartlett, S. J., Cheskin, L. J., & Pratt, M. (1998). Relationship of physical activity and television watching with body weight and level of fatness among children: Results from the Third National Health and Nutrition Examination Survey. *Jama*, *279*(12), 938–942.

Arias, E., MacDorman, M. F., Strobino, D. M., & Guyer, B. (2003). Annual summary of vital statistics—2002. *Pediatrics*, *112*(6), 1215–1230.

Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. *Journal of Psychiatric Research*, *33*(2), 97–104.

Arul, K., & Mesfin, A. (2017). The top 100 cited papers in health care disparities: A bibliometric analysis. *Journal of Racial and Ethnic Health Disparities*, *4*(5), 854–865.

Ashing-Giwa, K. T., Padilla, G., Tejero, J., Kraemer, J., Wright, K., Coscarelli, A., Clayton, S., Williams, I., & Hills, D. (2004). Understanding the breast cancer experience of women: A qualitative study of African American, Asian American, Latina and Caucasian cancer survivors. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, *13*(6), 408–428.

Atkins, R. (2014). Instruments measuring perceived racism/racial discrimination: Review and critique of factor analytic techniques. *International Journal of Health Services*, *44*(4), 711–734.

Baker, J. L., Rotimi, C. N., & Shriner, D. (2017). Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, *7*(1), 1572.

Banks, J., Marmot, M., Oldfield, Z., & Smith, J. P. (2006). Disease and disadvantage in the United States and in England. *Jama*, *295*(17), 2037–2045.

Barraza, F., Arancibia, M., Madrid, E., & Papuzinski, C. (2019). General concepts in biostatistics and clinical epidemiology: Random error and systematic error. *Medwave*, *19*(7), e7687.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2019). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation*, *18*(1), 6.

Bell, C. N., Thorpe, R. J., & LaVeist, T. A. (2018). The role of social context in racial disparities in self-rated health. *Journal of Urban Health*, *95*(1), 13–20.

Bhaskaran, K., & Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, *43*(4), 1336–1339.

Bhopal, R., & Donaldson, L. (1998). White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *American Journal of Public Health*, *88*(9), 1303–1307.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149.

Bobak, C. A., Barr, P. J., & O'Malley, A. J. (2018). Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology*, *18*(1), 1–11.

Bonham, V. L., & Nathan, V. R. (2002). Environmental public health research: Engaging communities. *International Journal of Hygiene and Environmental Health*, *205*(1–2), 11–18.

Brattain, M. (2007). Race, Racism, and Antiracism: UNESCO and the Politics of Presenting Science to the Postwar Public. *The American Historical Review*, *112*(5), 1386–1413. https://doi.org/10.1086/ahr.112.5.1386

Brown, A. (2020, February 25). The changing categories the U.S. census has used to measure race. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2020/02/25/the-changing-categories-the-u-s-has-used-to-measure-race/

Brown, T. N. (2003). Critical race theory speaks to the sociology of mental health: Mental health problems produced by racial stratification. *Journal of Health and Social Behavior*, *44*(3), 292–301.

Brownson, R. C., Jacobs, J. A., Tabak, R. G., Hoehner, C. M., & Stamatakis, K. A. (2013). Designing for dissemination among public health researchers: Findings from a national survey in the United States. *American Journal of Public Health*, *103*(9), 1693–1699.

Bryan, R. L., Kreuter, M. W., & Brownson, R. C. (2009). Integrating adult learning principles into training for public health practice. *Health Promotion Practice*, *10*(4), 557–563.

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: A simple guide for researchers. *The Malaysian Journal of Medical Sciences: MJMS*, *25*(6), 85.

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., Mountain, J. L., Pérez-Stable, E. J., Sheppard, D., & Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, *348*(12), 1170–1175.

Bureau, U. C. (n.d.). *2020 Census Illuminates Racial and Ethnic Composition of the Country*. Census.Gov. Retrieved August 8, 2022, from https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html

Bureau, U. C. (2022, March 1). *About the Topic of Race*. Census.Gov. https://www.census.gov/topics/population/race/about.html

Butler, J. 3rd, Fryer, C. S., Garza, M. A., Quinn, S. C., & Thomas, S. B. (2018). Commentary: Critical Race Theory Training to Eliminate Racial and Ethnic Health Disparities: The Public Health Critical Race Praxis Institute. *Ethnicity & Disease*, *28*(Suppl 1), 279–284. https://doi.org/10.18865/ed.28.S1.279

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio books.

Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, *9*, 403.

Castillo, W., & Gillborn, D. (2022). How to "QuantCrit:" Practices and questions for education data researchers and users. *EdWorkingPapers. Com*.

CDC. (1998, September 19). *Use of Race and Ethnicity in Public Health Surveillance Summary of the CDC/ATSDR Workshop*. https://www.cdc.gov/mmwr/preview/mmwrhtml/00021729.htm

CDC. (2021, November 24). *Racism and Health*. Centers for Disease Control and Prevention. https://www.cdc.gov/healthequity/racism-disparities/index.html

CDC. (2023, March 7). *To transform public health*. https://www.cdc.gov/minorityhealth/racism-disparities/expert-perspectives/besser/index.html

CDC - About BRFSS. (2019, February 9). https://www.cdc.gov/brfss/about/index.htm

Chambers, B. D., Baer, R. J., McLemore, M. R., & Jelliffe-Pawlowski, L. L. (2019). Using Index of Concentration at the Extremes as Indicators of Structural Racism to Evaluate the Association with Preterm Birth and Infant Mortality—California, 2011–2012. *Journal of Urban Health*, *96*(2), 159–170. https://doi.org/10.1007/s11524-018-0272-4

Choi, B. C., & Pak, A. W. (2005). Peer reviewed: A catalog of biases in questionnaires. *Preventing Chronic Disease*, *2*(1).

Christensen, L. B., & Waraczynski, M. A. (1988). *Experimental methodology*. Allyn and Bacon Boston.

Chu, J., Huang, W., Kuang, S., Wang, J., Xu, J., Chu, Z., Yang, Z., Lin, K., Li, P., & Wu, M. (1998). Genetic relationship of populations in China. *Proceedings of the National Academy of Sciences*, *95*(20), 11763–11768.

Churchwell, K., Elkind, M. S., Benjamin, R. M., Carson, A. P., Chang, E. K., Lawrence, W., Mills, A., Odom, T. M., Rodriguez, C. J., & Rodriguez, F. (2020). Call to action: Structural racism as a fundamental driver of health disparities: A presidential advisory from the American Heart Association. *Circulation*, *142*(24), e454–e468.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Cohn, D. (2010, January 21). Race and the Census: The "Negro" Controversy. *Pew Research Center's Social & Demographic Trends Project*. https://www.pewresearch.org/social-trends/2010/01/21/race-and-the-census-the-negro-controversy/

Cokley, K. (2007). Critical issues in the measurement of ethnic and racial identity: A referendum on the state of the field. *Journal of Counseling Psychology*, *54*(3), 224.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Houghton Mifflin Boston.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (Vol. 1195). Houghton Mifflin Boston, MA.

Cooper, R. S. (2013). Race in biological and biomedical research. *Cold Spring Harbor Perspectives in Medicine*, *3*(11), a008573.

Cordier, R., Speyer, R., Martinez, M., & Parsons, L. (2023). Reliability and Validity of Non-Instrumental Clinical Assessments for Adults with Oropharyngeal Dysphagia: A Systematic Review. *Journal of Clinical Medicine*, *12*(2), 721.

Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract. Assess. Res. Eval*, *10*, 1–10.

Country Comparisons—Population. (n.d.). Retrieved January 29, 2024, from https://www.cia.gov/the-world-factbook/field/population/country-comparison/

Crenshaw, K. (2015). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *The University of Chicago Legal Forum*, *1989*(1). WorldCat.org.

Cuevas, A. G., Chen, R., Slopen, N., Thurber, K. A., Wilson, N., Economos, C., & Williams, D. R. (2020). Assessing the role of health behaviors, socioeconomic status, and cumulative stress for racial/ethnic disparities in obesity. *Obesity*, *28*(1), 161–170.

Davis, D. W. (1997). Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans. *The Public Opinion Quarterly*, *61*(1), 183–207. JSTOR.

De Hert, M., Correll, C. U., Bobes, J., Cetkovich-Bakmas, M., Cohen, D., Asai, I., Detraux, J., Gautam, S., Möller, H.-J., & Ndetei, D. M. (2011). Physical illness in patients with severe mental

disorders. I. Prevalence, impact of medications and disparities in health care. *World Psychiatry*, *10*(1), 52.

de Winter, J. C., Dodou*, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147–181.

Devaraj, S., Stewart, A., Baumann, S., Elias, T. I., Hershey, T. B., Barinas-Mitchell, E., & Gary-Webb, T. L. (2020). Changes over time in disparities in health behaviors and outcomes by race in Allegheny County: 2009 to 2015. *Journal of Racial and Ethnic Health Disparities*, *7*(5), 838–843.

Doherty, E. E., & Green, K. M. (2023). Cohort Profile: The Woodlawn Study. *Journal of Developmental and Life-Course Criminology*, 1–24.

Downes, M. J., Brennan, M. L., Williams, H. C., & Dean, R. S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open*, *6*(12), e011458.

Dubay, L. C., & Lebrun, L. A. (2012). Health, behavior, and health care disparities: Disentangling the effects of income and race in the United States. *International Journal of Health Services*, *42*(4), 607–625.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

Dyer, L., Chambers, B. D., Crear-Perry, J., Theall, K. P., & Wallace, M. (2022). The Index of Concentration at the Extremes (ICE) and Pregnancy-Associated Mortality in Louisiana, 2016–2017. *Maternal and Child Health Journal*, *26*(4), 814–822. https://doi.org/10.1007/s10995-021-03189-1

Eigen, S., & Larrimore, M. (2012). *The german invention of race*. SUNY Press.

Eldridge, S., Campbell, M., Campbell, M., Dahota, A., Giraudeau, B., Higgins, J., Reeves, B., & Siegfried, N. (2016). Revised Cochrane risk of bias tool for randomized trials (RoB 2.0): Additional considerations for cluster-randomized trials. *Cochrane Methods. Cochrane Database Syst Rev*, *10*.

Faculty & Staff | University of Maryland | School of Public Health. (n.d.). Retrieved January 14, 2024, from https://sph.umd.edu/people/faculty-staff

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*(37), 15031–15036. https://doi.org/10.1073/pnas.1302997110

Feero, W. G., Steiner, R. D., Slavotinek, A., Faial, T., Bamshad, M. J., Austin, J., Korf, B. R., Flanagin, A., & Bibbins-Domingo, K. (2024). Guidance on use of race, ethnicity, and geographic origin as proxies for genetic ancestry groups in biomedical publications. *Human Genetics and Genomics Advances*.

Feldman, J. M., Waterman, P. D., Coull, B. A., & Krieger, N. (2015). Spatial social polarisation: Using the Index of Concentration at the Extremes jointly for income and race/ethnicity to analyse risk of hypertension. *Journal of Epidemiology and Community Health*, *69*(12), 1199. https://doi.org/10.1136/jech-2015-205728

Ferguson, L. (2004). External validity, generalizability, and knowledge utilization. *Journal of Nursing Scholarship*, *36*(1), 16–22.

Fernández-Férez, A., Ventura-Miranda, M. I., Camacho-Ávila, M., Fernández-Caballero, A., Granero-Molina, J., Fernández-Medina, I. M., & Requena-Mullor, M. del M. (2021). Nursing interventions to facilitate the grieving process after perinatal death: A systematic review. *International Journal of Environmental Research and Public Health*, *18*(11), 5587.

Fitriana, N., Hutagalung, F. D., Awang, Z., & Zaid, S. M. (2022). Happiness at work: A cross-cultural validation of happiness at work scale. *Plos One*, *17*(1), e0261617.

Flanagin, A., Frey, T., Christiansen, S. L., & AMA Manual of Style Committee. (2021). Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA*, *326*(7), 621–627. https://doi.org/10.1001/jama.2021.13304

Flanagin, A., Frey, T., Christiansen, S. L., & Bauchner, H. (2021). The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. *JAMA*, *325*(11), 1049–1052. https://doi.org/10.1001/jama.2021.2104

Flannelly, K. J., Flannelly, L. T., & Jankowski, K. R. B. (2018). Threats to the Internal Validity of Experimental and Quasi-Experimental Research in Healthcare. *Journal of Health Care Chaplaincy*, *24*(3), 107–130. https://doi.org/10.1080/08854726.2017.1421019

Forbes, R. P. (2009). *The Missouri compromise and its aftermath: Slavery and the meaning of America*. Univ of North Carolina Press.

Ford, C. L., & Airhihenbuwa, C. O. (2018). Commentary: Just What is Critical Race Theory and What's it Doing in a Progressive Field like Public Health? *Ethnicity & Disease*, *28*(Suppl 1), 223–230. https://doi.org/10.18865/ed.28.S1.223

Ford, C. L., Ford, C. L., & Airhihenbuwa, C. O. (2010). *Critical Race Theory, Race Equity, and Public Health: Toward Antiracism Praxis*. WorldCat.org. http://hdl.handle.net/1903/23394

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Sage.

Franco, O. H., Calkins, M. E., Giorgi, S., Ungar, L. H., Gur, R. E., Kohler, C. G., & Tang, S. X. (2022). Feasibility of Mobile Health and Social Media-Based Interventions for Young Adults With Early Psychosis and Clinical Risk for Psychosis: Survey Study. *JMIR Formative Research*, *6*(7), e30230. cmedm. https://doi.org/10.2196/30230

Frank, J., Marion, G., & Doeschl-Wilson, A. (2021). Development of a critical appraisal tool for models predicting the impact of 'test, trace, and protect'programmes on COVID-19 transmission. *Public Health*, *201*, 55–60.

Frank, R. (2007). What to make of it? The (Re) emergence of a biological conceptualization of race in health disparities research. *Social Science & Medicine*, *64*(10), 1977–1983.

Freeman, H. P. (1998). The meaning of race in science--considerations for cancer research: Concerns of special populations in the National Cancer Program. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *82*(1), 219–225.

Fullilove, M. T. (1998). Comment: Abandoning" race" as a variable in public health research—An idea whose time has come. *American Journal of Public Health*, *88*(9), 1297–1298.

Gannon, M. (2016). Race is a social construct, scientists argue. *Scientific American*, *5*, 1–11.

Gao, Z. (2020). Researcher biases. *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*, 37–41.

Garcia, J. A., Sanchez, G. R., Sanchez-Youngman, S., Vargas, E. D., & Ybarra, V. D. (2015). RACE AS LIVED EXPERIENCE: The Impact of Multi-Dimensional Measures of Race/Ethnicity on the Self-Reported Health Status of Latinos. *Du Bois Review: Social Science Research on Race*, *12*(2), 349–373. Cambridge Core. https://doi.org/10.1017/S1742058X15000120

Garvey, A., Lynch, G., Mansour, M., Coyle, A., Gard, S., & Truglio, J. (2022). From Race to Racism: Teaching a Tool to Critically Appraise the Use of Race in Medical Research. *MedEdPORTAL : The Journal of Teaching and Learning Resources*, *18*, 11210. cmedm. https://doi.org/10.15766/mep_2374-8265.11210

Gee, G. C., & Ford, C. L. (2011). Structural racism and health inequities: Old issues, New Directions1. *Du Bois Review: Social Science Research on Race*, *8*(1), 115–132.

Gibbons, J., & Barton, M. S. (2016). The Association of Minority Self-Rated Health with Black versus White Gentrification. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, *93*(6), 909–922. WorldCat.org. https://doi.org/10.1007/s11524-016-0087-0

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338.

Gokmen, S., Dagalp, R., & Kilickaplan, S. (2022). Multicollinearity in measurement error models. *Communications in Statistics-Theory and Methods*, *51*(2), 474–485.

Gomer, B. (2019). MCAR, MAR, and MNAR values in the same dataset: A realistic evaluation of methods for handling missing data. *Multivariate Behavioral Research*, *54*(1), 153–153.

Goodwin, L. D., & Goodwin, W. L. (1991). Focus on psychometrics. Estimating construct validity. *Research in Nursing & Health*, *14*(3), 235–243.

google random number generator—Google Search. (n.d.). Retrieved July 30, 2023, from https://www.google.com/search?si=ACFMAn9VGkye4lX6FGB1mT7l9DMe3bukTuLQ_mzZDiYKqOzy4Ll4ax4kKH0st9XiBjZYnfFIKeUhWsaLZXLhlaLCfv5Zj62f-A5J2jdj0xAACKD3w4U95aA%3D&hl=en-US&kgs=33fa213409437ebf&shndl=21&source=sh/x/fbx/m1/1

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hahn, R. A., Truman, B. I., & Barker, N. D. (1996). Identifying ancestry: The reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. *Epidemiology*, 75–80.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.

Haq, K. S., & Penning, M. J. (2020). Social determinants of racial disparities in cognitive functioning in later life in Canada. *Journal of Aging and Health*, *32*(7–8), 817–829.

Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., & Wang, X. (2018). Use of reproducible research practices in public health: A survey of public health analysts. *PLoS One*, *13*(9), e0202447.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing*, *18*(3), 66. https://doi.org/10.1136/eb-2015-102129

Heise, T. L., Seidler, A., Girbig, M., Freiberg, A., Alayli, A., Fischer, M., Haß, W., & Zeeb, H. (2022). CAT HPPR: a critical appraisal tool to assess the quality of systematic, rapid, and scoping

reviews investigating interventions in health promotion and prevention. *BMC Medical Research Methodology*, *22*(1), 334.

HeyGen—AI Video Generator. (n.d.). Retrieved February 1, 2024, from https://www.heygen.com/

Hochschild, J. L., & Powell, B. M. (2008). Racial reorganization and the United States Census 1850–1930: Mulattoes, half-breeds, mixed parentage, Hindoos, and the Mexican race. *Studies in American Political Development*, *22*(1), 59–96.

Home. (n.d.). Canva. Retrieved February 1, 2024, from https://www.canva.com/

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., & O'Cathain, A. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291.

Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). *Journal of Evaluation in Clinical Practice*, *24*(3), 459–467.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., & Nicolau, B. (2019). Improving the content validity of the mixed methods appraisal tool: A modified e-Delphi study. *Journal of Clinical Epidemiology*, *111*, 49–59.

Hsu, P., Bryant, M. C., Hayes-Bautista, T. M., Partlow, K. R., & Hayes-Bautista, D. E. (2019). Racially Ambiguous Babies and Racial Narratives in the United States: A Growing Contradiction for Health Disparities Research. *Academic Medicine : Journal of the Association of American Medical Colleges*, *94*(8), 1099–1102. cmedm. https://doi.org/10.1097/ACM.0000000000002740

IAM, I. (2020). *Acceptability of Intervention Measure (AIM), Intervention Appropriateness Measure (IAM), & Feasibility of Intervention Measure. Ictp. Fpg. Unc. Edu*.

In, J. (2017). Introduction of a pilot study. *Korean Journal of Anesthesiology*, *70*(6), 601–605.

Ioannidis, J. P., Powe, N. R., & Yancy, C. (2021). Recalibrating the use of race in medical research. *Jama*, *325*(7), 623–624.

Jackson, K. F. (2012). Living the multiracial experience: Shifting racial expressions, resisting race, and seeking community. *Qualitative Social Work*, *11*(1), 42–60.

JBI Critical Appraisal Tools | JBI. (n.d.). Retrieved January 23, 2024, from https://jbi.global/critical-appraisal-tools

Jeanfreau, S. G., & Jack Jr, L. (2010). Appraising qualitative research in health education: Guidelines for public health educators. *Health Promotion Practice*, *11*(5), 612–617.

Jeyaraman, M. M., Al-Yousif, N., Robson, R. C., Copstein, L., Balijepalli, C., Hofer, K., Fazeli, M. S., Ansari, M. T., Tricco, A. C., & Rabbani, R. (2020). Inter-rater reliability and validity of risk of bias instrument for non-randomized studies of exposures: A study protocol. *Systematic Reviews*, *9*(1), 1–12.

Jin, M., Liu, A., Chen, Z., & Li, Z. (2013). Sequential testing of measurement errors in inter-rater reliability studies. *Statistica Sinica*, *23*(4), 1743.

Johnson, V. E., & Carter, R. T. (2020). Black Cultural Strengths and Psychosocial Well-Being: An Empirical Analysis With Black American Adults. *Journal of Black Psychology*, *46*(1), 55–89. https://doi.org/10.1177/0095798419889752

Jones, C. P. (2000). Levels of racism: A theoretic framework and a gardener's tale. *American Journal of Public Health*, *90*(8), 1212.

Jones, C. P. (2001). Invited commentary:"race," racism, and the practice of epidemiology. *American Journal of Epidemiology*, *154*(4), 299–304.

Jones, D. S., Hammonds, E., Gone, J. P., & Williams, D. (2024). Explaining Health Inequities—The Enduring Legacy of Historical Biases. *New England Journal of Medicine*.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36.

Kao, G., & Vaquera, E. (2006). The salience of racial and ethnic identification in friendship choices among Hispanic adolescents. *Hispanic Journal of Behavioral Sciences*, *28*(1), 23–47.

Kaufman, J. S., & Cooper, R. S. (2001). Commentary: Considerations for Use of Racial/Ethnic Classification in Etiologic Research. *American Journal of Epidemiology*, *154*(4), 291–298. https://doi.org/10.1093/aje/154.4.291

Khullar, D., & Chokshi, D. A. (2018). Health, income, & poverty: Where we are & what could help. *Health Affairs*, *10*(10.1377).

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233.

Kittles, R. A., Santos, E. R., Oji-Njideka, N. S., & Bonilla, C. (2007). Race, Skin Color and Genetic Ancestry. *Californian Journal of Health Promotion*, *5*(SI), 9–23.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kmet, L. M., Cook, L. S., & Lee, R. C. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*.

Knowles, M. S. (1980). From pedagogy to andragogy. *Religious Education*.

Knowles, M. S. (1984). *Theory of andragogy*.

Knowles, M. S., Holton III, E. F., & Swanson, R. A. (2014). *The adult learner: The definitive classic in adult education and human resource development*. Routledge.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Ladson-Billings, G., & Tate, W. F. (1995). Toward a critical race theory of education. *Teachers College Record*, *97*(1), 47–68.

Lang, J. M., Rothman, K. J., & Cann, C. I. (1998). That confounded P-value. *Epidemiology (Cambridge, Mass.)*, *9*(1), 7–8.

Larsson, H., Tegern, M., Monnier, A., Skoglund, J., Helander, C., Persson, E., Malm, C., Broman, L., & Aasa, U. (2015). Content validity index and intra-and inter-rater reliability of a new muscle strength/endurance test battery for Swedish soldiers. *PloS One*, *10*(7), e0132185.

LaVeist, T. A. (1994). Beyond dummy variables and sample selection: What health services researchers ought to know about race as a variable. *Health Services Research*, *29*(1), 1.

LaVeist, T. A. (2000). On the study of race, racism, and health: A shift from description to explanation. *International Journal of Health Services*, *30*(1), 217–219.

LaVeist, T. A. (2005). *Minority populations and health: An introduction to health disparities in the United States* (Vol. 4). John Wiley & Sons.

LaVeist, T., Pollack, K., Thorpe Jr, R., Fesahazion, R., & Gaskin, D. (2011). Place, not race: Disparities dissipate in southwest Baltimore when blacks and whites live under similar conditions. *Health Affairs*, *30*(10), 1880–1887.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Sage publications.

Lawrence, K., & Keleher, T. (2004, November 11). *Chronic disparity: Strong and pervasive evidence of racial inequalities.* Race and Public Policy Conference, Berkeley, CA.

Leary, H., Giersch, S., Walker, A., & Recker, M. (2009). *Developing a review rubric for learning resources in digital libraries*. 421–422.

Leary, H., Giersch, S., Walker, A., & Recker, M. (2011). Developing and using a guide to assess learning resource quality in educational digital libraries. *Digital Libraries-Methods and Applications*, 181–196.

Lee, T. A., & Pickard, A. S. (2013). Exposure Definition and Measurement. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Agency for Healthcare Research and Quality (US). http://www.ncbi.nlm.nih.gov/books/NBK126191/

Liang, K.-Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, *14*(1), 43–68.

Lin, S. S., & Kelsey, J. L. (2000). Use of race and ethnicity in epidemiologic research: Concepts, methodological issues, and suggestions for research. *Epidemiologic Reviews*, *22*(2), 187–202.

Lund, H., Robinson, K. A., Gjerland, A., Nykvist, H., Drachen, T. M., Christensen, R., Juhl, C. B., Jamtvedt, G., Nortvedt, M., & Bjerrum, M. (2022). Meta-research evaluating redundancy and use

of systematic reviews when planning new studies in health research: A scoping review. *Systematic Reviews*, *11*(1), 241.

Luo, J., Hendryx, M., & Wang, F. (2022). Mortality disparities between Black and White Americans mediated by income and health behaviors. *SSM-Population Health*, *17*, 101019.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*.

Mack, C., Su, Z., & Westreich, D. (2018). Types of Missing Data. In *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet]*. Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK493614/

Martinez, R. A. M., Andrabi, N., Goodwin, A. N., Wilbur, R. E., Smith, N. R., & Zivich, P. N. (2022). Conceptualization, Operationalization, and Utilization of Race and Ethnicity in Major Epidemiology Journals 1995–2018: A Systematic Review. *American Journal of Epidemiology*, kwac146. https://doi.org/10.1093/aje/kwac146

Mateen, F. J., Oh, J., Tergas, A. I., Bhayani, N. H., & Kamdar, B. B. (2013). Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clinical Epidemiology*, 89–95.

Mateo, C. M., & Williams, D. R. (2021). Racism: A fundamental driver of racial disparities in health-care quality. *Nature Reviews Disease Primers*, *7*(1), 20. https://doi.org/10.1038/s41572-021-00258-1

Matthay, E. C., & Glymour, M. M. (2020). A graphical catalog of threats to validity: Linking social science with epidemiology. *Epidemiology (Cambridge, Mass.)*, *31*(3), 376.

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*, *50*(5), 504–519.

McClendon, J., Chang, K., Boudreaux, M. J., Oltmanns, T. F., & Bogdan, R. (2021). Black-White racial health disparities in inflammation and physical health: Cumulative stress, social isolation, and health behaviors. *Psychoneuroendocrinology*, *131*, 105251.

McDermott, R. (2011). Internal and external validity. *Cambridge Handbook of Experimental Political Science*, 27–40.

McIntosh, K., Moss, E., Nunn, R., & Shambaugh, J. (2020). Examining the Black-white wealth gap. *Washington DC: Brooking Institutes*.

Meissner, H. I., Breen, N., Klabunde, C. N., & Vernon, S. W. (2006). Patterns of colorectal cancer screening uptake among men and women in the United States. *Cancer Epidemiology Biomarkers & Prevention*, *15*(2), 389–394.

Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2005). State of disparities in cardiovascular health in the United States. *Circulation*, *111*(10), 1233–1241.

Mezuk, B., Rafferty, J. A., Kershaw, K. N., Hudson, D., Abdou, C. M., Lee, H., Eaton, W. W., & Jackson, J. S. (2010). Reconsidering the role of social disadvantage in physical and mental health: Stressful life events, health behaviors, race, and depression. *American Journal of Epidemiology*, *172*(11), 1238–1249.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308–2316.

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A., & Ferreira, M. E. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, *30*.

Morgenstern, H. (1997). Defining and Explaining Race Effects. *Epidemiology*, *8*(6), 609–611. JSTOR.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168.

Munn, Z., Barker, T. H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M., & Aromataris, E. (2020). Methodological quality of case series studies: An introduction to the JBI critical appraisal tool. *JBI Evidence Synthesis*, *18*(10), 2127–2133.

Munn, Z., Moola, S., Riitano, D., & Lisy, K. (2014). The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. *International Journal of Health Policy and Management*, *3*(3), 123.

Murray, C. J., Kulkarni, S., & Ezzati, M. (2005). Eight Americas: New perspectives on US health disparities. *American Journal of Preventive Medicine*, *29*(5), 4–10.

Murray Horwitz, M. E., Pace, L. E., & Ross-Degnan, D. (2018). Trends and disparities in sexual and reproductive health behaviors and service use among young adult women (aged 18–25 years) in the United States, 2002–2015. *American Journal of Public Health*, *108*(S4), S336–S343.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20.

National Academies of Sciences, E., and Medicine. (2023). *Using population descriptors in genetics and genomics research: A new framework for an evolving field*.

National Academies of Sciences, Engineering, and Medicine. (n.d.). *The Use of Race and Ethnicity in Biomedical Research*. Retrieved February 1, 2024, from https://www.nationalacademies.org/our-work/the-use-of-race-and-ethnicity-in-biomedical-research#sectionWebFriendly

NIH stands against structural racism in biomedical research. (2021, February 26). National Institutes of Health (NIH). https://www.nih.gov/about-nih/who-we-are/nih-director/statements/nih-stands-against-structural-racism-biomedical-research

Noble, D. W., Lagisz, M., O'dea, R. E., & Nakagawa, S. (2017). *Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses*.

Nobles, M. (2000). *Shades of citizenship: Race and the census in modern politics*. Stanford University Press.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1609406919899220.

Onwuegbuzie, A. J. (2000). *Expanding the framework of internal and external validity in quantitative research.*

Onwuegbuzie, A. J., & Daniel, L. G. (2003). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education*, *6*.

Otten, M. W., Teutsch, S. M., Williamson, D. F., & Marks, J. S. (1990). The effect of known risk factors on the excess mortality of black adults in the United States. *Jama*, *263*(6), 845–850.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., & Seller, R. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies*, *49*(1), 47–53.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 1–11.

Parcha, V., Malla, G., Suri, S. S., Kalra, R., Heindl, B., Berra, L., Fouad, M. N., Arora, G., & Arora, P. (2020). Geographic variation in racial disparities in health and coronavirus Disease-2019 (COVID-19) mortality. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, *4*(6), 703–716.

Park, S., & Chung, Y. (2022). The effect of missing levels of nesting in multilevel analysis. *Genomics & Informatics*, *20*(3).

Parker, L. J., Gaugler, J. E., & Gitlin, L. N. (2022). Use of Critical Race Theory to Inform the Recruitment of Black/African American Alzheimer's Disease Caregivers into Community-Based Research. *The Gerontologist*, *62*(5), 742–750. https://doi.org/10.1093/geront/gnac001

Perplexity. (n.d.). Retrieved January 17, 2024, from https://www.perplexity.ai/

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112.

Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*, *13*(1), 1–14.

Pierre, J. (2020). Slavery, Anthropological Knowledge, and the Racialization of Africans. *Current Anthropology*, *61*(S22), S220–S231. https://doi.org/10.1086/709844

Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. *International Journal of Nursing Studies*, *46*(4), 529–546.

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health*, *35*, 29–45.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Health, P., i, M., & Research, MHS (2011). Outcomes for implementation research: Conceptual distinctions. *Measurement Challenges, and Research Agenda*, *38*(2), 65–76.

Redmond, N., Baer, H. J., & Hicks, L. S. (2011). Health behaviors and racial disparity in blood pressure control in the national health and nutrition examination survey. *Hypertension*, *57*(3), 383–389.

Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. (2024, March 29). Federal Register. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and

Riley, R. D., Cole, T. J., Deeks, J., Kirkham, J. J., Morris, J., Perera, R., Wade, A., & Collins, G. S. (2022). On the 12th Day of Christmas, a Statistician Sent to Me... *Bmj*, *379*.

Rodrigues, I. B., Adachi, J. D., Beattie, K. A., & MacDermid, J. C. (2017). Development and validation of a new tool to measure the facilitators, barriers and preferences to exercise in people with osteoporosis. *BMC Musculoskeletal Disorders*, *18*(1), 1–9.

Root, M. (2008). Measurement error in racial and ethnic statistics. *Biology & Philosophy*, *24*(3), 375. https://doi.org/10.1007/s10539-008-9138-6

Ross, P. T., Hart-Johnson, T., Santen, S. A., & Zaidi, N. L. B. (2020). Considerations for using race and ethnicity as quantitative variables in medical education research. *Perspectives on Medical Education*, *9*(5), 318–323.

Rutkowski, L., Svetina, D., & Liaw, Y.-L. (2019). Collapsing Categorical Variables and Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 790–802. https://doi.org/10.1080/10705511.2018.1547640

Saperstein, A., & Penner, A. M. (2012). Racial fluidity and inequality in the United States. *American Journal of Sociology*, *118*(3), 676–727.

Schmidt, H., Gostin, L. O., & Williams, M. A. (2023). The Supreme Court's Rulings on Race Neutrality Threaten Progress in Medicine and Health. *JAMA*.

Shavers, V. L., & Brown, M. L. (2002). Racial and ethnic disparities in the receipt of cancer treatment. *Journal of the National Cancer Institute*, *94*(5), 334–357.

Skinner, J., Weinstein, J. N., Sporer, S. M., & Wennberg, J. E. (2003). Racial, ethnic, and geographic disparities in rates of knee arthroplasty among Medicare patients. *New England Journal of Medicine*, *349*(14), 1350–1359.

Slack, M. K., & Draugalis, J. R., Jr. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy*, *58*(22), 2173–2181. https://doi.org/10.1093/ajhp/58.22.2173

Smith, R. J., Pride, T. T., & Schmitt-Sands, C. E. (2017). Does spatial assimilation lead to reproduction of gentrification in the global city? *Journal of Urban Affairs*, *39*(6), 745–763. WorldCat.org. https://doi.org/10.1080/07352166.2016.1262693

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care*, 733–741.

Speyer, R., Denman, D., Wilkes-Gillan, S., Chen, Y., Bogaardt, H., Kim, J., Heckathorn, D., & Cordier, R. (2018). Effects of telehealth by allied health professionals and nurses in rural and remote

areas: A systematic review and meta-analysis. *Journal of Rehabilitation Medicine*, *50*(3), 225–235.

Srivastav, A., Robinson-Ector, K., Kipp, C., Strompolis, M., & White, K. (2021). Who declines to respond to the reactions to race module?: Findings from the South Carolina Behavioral Risk Factor Surveillance System, 2016–2017. *BMC Public Health*, *21*, 1–12.

Statistical Brief on the Reactions to Race Module, Behavioral Risk Factor Surveillance System, 2022. (n.d.).

Stovall, D. (2005). A challenge to traditional theory: Critical race theory, African-American community organizers, and education. *Discourse*, *26*(1), 95–108. WorldCat.org.

Strömberg, U. (1996). Collapsing ordered outcome categories: A note of concern. *American Journal of Epidemiology*, *144*(4), 421–424.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., & Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Jama*, *283*(15), 2008–2012.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). pearson Boston, MA.

Tajfel, H., & Turner, J. (1979). „An integrative theory of inter-group conflict ", Austin, W.─ S. *The Social Psychology of Inter-Group Relations. Monterey, CA: Brooks/Cole*.

Thomas, B., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing*, *1*(3), 176–184.

Thorpe Jr, R. J., Kennedy-Hendricks, A., Griffith, D. M., Bruce, M. A., Coa, K., Bell, C. N., Young, J., Bowie, J. V., & LaVeist, T. A. (2015). Race, social and environmental conditions, and health behaviors in men. *Family & Community Health*, *38*(4), 297.

Tran, V. T., & Johnston-Guerrero, M. P. (2016). Is transracial the same as transgender? The utility and limitations of identity analogies in multicultural education. *Multicultural Perspectives*, *18*(3), 134–139.

Trivedi, A. N., Zaslavsky, A. M., Schneider, E. C., & Ayanian, J. Z. (2005). Trends in the quality of care and racial disparities in Medicare managed care. *New England Journal of Medicine*, *353*(7), 692–700.

Trochim, W., & Donnelly, J. (2006). The research methods knowledge base. 3rd. *Mason, OH: Atomic Dog Publishing*.

Trochim, W. M. K. (n.d.-a). *Knowledge Base*. Retrieved January 17, 2024, from https://conjointly.com/kb/

Trochim, W. M. K. (n.d.-b). *Nonprobability Sampling*. Retrieved January 17, 2024, from https://conjointly.com/kb/nonprobability-sampling/

Troncoso Skidmore, S., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, *45*(2), 536–546.

Tulier ME, Reid C, Mujahid MS, & Allen AM. (2019). "Clear action requires clear thinking": A systematic review of gentrification and health research in the United States. *Health & Place, 59*, 102173. WorldCat.org. https://doi.org/10.1016/j.healthplace.2019.102173

Tuthill, Z., Denney, J. T., & Gorman, B. (2020). Racial disparities in health and health behaviors among gay, lesbian, bisexual and heterosexual men and women in the BRFSS-SOP. *Ethnicity & Health*, *25*(2), 177–188. https://doi.org/10.1080/13557858.2017.1414157

United States—Census Bureau Profile. (n.d.). Retrieved January 28, 2024, from https://data.census.gov/profile/United_States?g=010XX00US

U.S. Census Bureau. (2021). *Hispanic or Latino Origin by Race: 2010 and 2020*. U.S. Census Bureau,. https://www2.census.gov/programs-surveys/decennial/2020/data/redistricting-supplementary-tables/redistricting-supplementary-table-04.pdf

US Census Bureau Regions and Divisions with State FIPS Codes. (n.d.). Retrieved November 23, 2023, from https://www2.census.gov/geo/docs/maps-data/maps/reg_div.txt

US changes how it categorizes people by race and ethnicity. It's the first revision in 27 years. (2024, March 28). AP News. https://apnews.com/article/race-ethnicity-census-bureau-hispanics-0b2c325b683efd95e8e8e24235654abd

Van Doorslaer, E., Koolman, X., & Jones, A. M. (2002). Explaining income-related inequalities in doctor utilisation in Europe: A decomposition approach. *ECuity II Project*.

Vargas, N. (2015). LATINA/O WHITENING?: Which Latina/os self-classify as White and report being perceived as White by other Americans? *Du Bois Review: Social Science Research on Race*, *12*(1), 119–136.

Viswanathan, M. (2005). *Measurement error and research design*. Sage.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, *5*(4), 425.

Wang, H. L. (2018, February 1). 2020 Census Will Ask White People More About Their Ethnicities. *NPR*. https://www.npr.org/2018/02/01/582338628/-what-kind-of-white-2020-census-to-ask-white-people-about-origins

Wang, H. L. (2022, June 15). Biden officials may change how the U.S. defines racial and ethnic groups by 2024. *NPR*. https://www.npr.org/2022/06/15/1105104863/racial-ethnic-categories-omb-directive-15

Wang, H. L. (2024, March 28). Next U.S. census will have new boxes for "Middle Eastern or North African," "Latino." *NPR*. https://www.npr.org/2024/03/28/1237218459/census-race-categories-ethnicity-middle-east-north-africa

Waters, M. C. (2000). Immigration, intermarriage, and the challenges of measuring racial/ethnic identities. *American Journal of Public Health*, *90*(11), 1735.

Wedow, R., Martschenko, D. O., & Trejo, S. (2022). Scientists must consider the risk of racist misappropriation of research. *Scientific American*.

Weiner, B. J., Lewis, C. C., Stanick, C., Powell, B. J., Dorsey, C. N., Clary, A. S., Boynton, M. H., & Halko, H. (2017). Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science*, *12*(1), 1–12.

West, C. (1995). *Critical race theory: The key writings that formed the movement*. The New Press.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608.

Whitaker, K. M., Jacobs Jr, D. R., Kershaw, K. N., Demmer, R. T., Booth III, J. N., Carson, A. P., Lewis, C. E., Goff Jr, D. C., Lloyd-Jones, D. M., & Gordon-Larsen, P. (2018). Racial disparities in cardiovascular health behaviors: The coronary artery risk development in young adults study. *American Journal of Preventive Medicine*, *55*(1), 63–71.

White, M. (2022). Sample size in quantitative instrument validation studies: A systematic review of articles published in Scopus, 2021. *Heliyon*, *8*(12).

Williams, C., Birungi, J., Brown, M., Deutsch, J., Williams, F., Perkins, P., Bishop, P., Walker, D., Moody, E., & Hamilton, R. (2022). Public Health Liberation–An Emerging Transdiscipline to Elucidate and Transform the Public Health Economy. *Advances in Clinical Medical Research and Healthcare Delivery*, *2*(3), 10.

Williams, C., Woodard, N., & Chao-Li Kuo, C. (2022). Reliability, Validity, and Exploratory Factor Analyses of Gentrification Health Research Measures. *Advances in Clinical Medical Research and Healthcare Delivery*, *2*(4), 1.

Williams, D. R., & Mohammed, S. A. (2013). Racism and health I: Pathways and scientific evidence. *American Behavioral Scientist*, *57*(8), 1152–1173.

Witzig, R. (1996). The Medicalization of Race: Scientific Legitimization of a Flawed Social Construct. *Annals of Internal Medicine*, *125*(8), 675–679. https://doi.org/10.7326/0003-4819-125-8-199610150-00008

Xavier, C., Benoit, A., & Brown, H. K. (2018). Teenage pregnancy and mental health beyond the postpartum period: A systematic review. *J Epidemiol Community Health*, *72*(6), 451–457.

Xiao, H., Vaidya, R., Liu, F., Chang, X., Xia, X., & Unger, J. M. (2023). Sex, racial, and ethnic representation in COVID-19 clinical trials: A systematic review and meta-analysis. *JAMA Internal Medicine*.

Yuan, M., & Recker, M. (2015). Not all rubrics are equal: A review of rubrics for evaluating the quality of open educational resources. *International Review of Research in Open and Distributed Learning*, *16*(5), 16–38.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4*(2), 165.