CONSTRUCT(ION) AND CONTEXT A Response to Methodological Issues in Studying Character

Nancy L. Deutsch
University of Virginia, Curry School of Education

In this article, I respond to Noel Card's *Methodological Issues in Measuring the Development of Character*. I focus on the ways in which social scientific knowledge represents human constructions of the world and the implications of this stance for the measurement of character. Further, I consider how context influences those constructions and the need for researchers to more purposefully engage with questions of construct (in)stability across contexts both within and between people. Finally, I discuss the ways in which expanding our methodological tool box, to include multiple quantitative and qualitative methods, will help us gain a more nuanced, and thereby more broadly useful, understanding of character development.

As I read Noel Card's article on methodological issues in studying character development (Card, this volume), two words kept popping out at me: construct and context. This might not be surprising. I am a qualitative and mixed methods inclined researcher. Studying how people construct meaning of and in their lives and how the social contexts in which they do so influence those meanings and, thereby, their behavior, are integral parts of what I study. That is not to say that quantitatively inclined researchers are not interested in these things. Rather, qualitative and quantitative researchers tend to approach constructs and contexts in different ways that make them more or less central, or at least central in different ways, to the questions that we ask about the world. And that is appropriate. Our questions drive our methods, and quantitative and qualitative methods have different, and complementary, strengths in terms of the types of knowledge and understanding that they address (see Yoshikawa, Weisner, Kalil, & Way, 2008, for discussion of the importance of both numeric and nonnumeric ways of understanding in developmental research).

Quantitative research is well suited for assessing the presence, amount, and prevalence of constructs related to character and to testing hypotheses about relationships between

• Correspondence concerning this article should be addressed to: Nancy L. Deutsch, nancyd@virginia.edu

Journal of Character Education, Volume 13(2), 2017, pp. 53–63 ISSN 1543-1223

character and other constructs or actions. Card's article does a fine job at pointing out the methodological issues that arise when considering those features of character. Yet there are other elements of character development that are left unaddressed with such methods, as well as other questions that are raised. Card's article hints at this when he discusses random effects models and how they acknowledge that some of the variability in statistical models is not statistical error, but is, as he writes, "variability that is real." It is this "real" variability, and its implications for how we measure character development, on which I want to focus. I will do so by digging into the terms construct and context and how I think they should inform our methods.

What struck me in Card's article is how the implications of context for our methods differ widely depending on how we define the word "construct." Which brings me to why the word construct struck me so powerfully. In developmental research, we continually refer to constructs. We measure them. We assess their prevalence and levels. We consider different conceptualizations of them, and the effects of those conceptualizations on what we conclude about the constructs. We study the relationship between constructs and test interventions that we think will have an effect on a construct, like character, that we believe, or that prior research has suggested, is important for human development. Yet, seldom in our day-to-day research or program lives do we deeply interrogate the basic assumptions we make about the constructs we measure. Seldom do we acknowledge that all *constructs* are *social constructions* (see Berger & Luckmann, 1966; Gergen, 1985; Guba & Lincoln, 1994; and Latour, 1991, for discussions of social and human constructions of reality, psychological constructs, paradigms of social research, and science, respectively). And so, it is the nature of character as a social construction, the ways in which the very construction of character as a construct may differ across contexts, and the implications of these for how we study character that I address. I acknowledge the tension between seeking consistency of constructs and recognizing that context matters, and attempt to emphasize how that tension can improve the field of character development rather than stymie it.

When we talk about construct validity or measurement reliability, we are acknowledging that the instruments we use to assess a construct like character are human constructions, and thus subject to error. We also recognize that there can be multiple measures that all claim to study character, and that do so with equivalent levels of validity and reliability, but that focus on different aspects of character. This is what Figure 1 from Card's article illustrates. Yet when we examine measures in this way we are still assuming that something called character exists in the world as a singular, objective fact. The issue we grapple with in measurement, then, is not whether character exists, but how close we are able to get to the "real" or "true" (and I use quotation marks intentionally) construct of character. But all constructs in social science are human constructions and, as such, are bound by our social, historical, and political times and lenses (see Berger & Luckmann, 1966; Gergen, 1985; Guba & Lincoln, 1994; Latour, 1991).

Take psychopathology as an example. Up until 1973, homosexuality was defined as a sociopathic personality disorder in the American Psychiatric Association's Diagnostics and Statistical Manual of Mental Disorders (DSM; see Karasic & Drescher, 2006 for review of changes in the DSM of diagnoses related to sexuality and gender). Thus, the construct of homosexuality included mental illness as one of its core defining features. Over the past 40 years, the construct of homosexuality has evolved. In 2009, the American Psychological Association released a resolution making clear that homosexuality is a "normal and positive variation of human sexuality" (Anton, 2010, p. 29) and opposing the use of conversion therapies (Anton, 2010). And today the construct of homosexuality is still being redefined. The very terms we use today, LGBTQ being just one example, reflect the still broadening defi-

nition of that early construct in accordance with people's lived realities. So, constructs change.

Yet how we define, and therefore measure, a construct at any given point in time has implications for what and who we consider "high," "low," or "normative" in that domain, and therefore how and with whom we intervene to affect change in that domain. Certainly, the change in stance towards conversion therapies is one example. And the field of character development has not been immune to such shifts. Other authors in this volume and elsewhere have noted the breadth and variety of definitions and terms used within the study of character development (see, i.e., Seider, Jayawickreme, & Lerner, 2017), one indication of changing constructs. Most definitions of character, however, rely on a foundation of moral development. For example, Berkowitz and Bier (2014) define character as "the set of psychological characteristics that motivate and enable an individual to function as a competent moral agent" (p. 250). Yet the nature of morality, a foundational component of character by this and other definitions, has itself been the subject of debate.

In the late 1950s, Lawrence Kohlberg began studying moral development in samples of White men. From this research, he developed six stages of moral development through which he proposed that people progress, with each stage representing a higher level of moral reasoning (Kohlberg, 1976, 1981). Since the measure's initial development, the stages have largely held up across different cultural contexts, with some differences in the *content* of postconventional reasoning but stability in the stages themselves (Snarey & Samuelson, 2014). Yet the stability of the construct across gender provoked controversy in the field. In 1982, Carol Gilligan published *In a Different Voice*, in which she suggested that Kohlberg's theory did not adequately represent the moral reasoning and development of women. Rather than studying moral development using hypothetical scenarios, as Kohlberg did, Gilligan studied women who faced a real-life moral dilemma. In doing so, she concluded that Kohlberg's results were "sex-biased against a moral orientation based on care, in contrast to his positing of a justice orientation as the goal for mature moral development" (Larrabee, 1993, p. 4). Over the next few decades, researchers continued to test this theory empirically (see Baumrind, 1986, and Walker, 1984, reprinted in Larrabee, 1993, for early reviews of and responses to these empirical tests). And Gilligan herself was critiqued both by Kohlberg's supporters and by feminists, who expressed concern over the lack of racial and ethnic diversity in her sample as well as apprehension over the ways in which her work could reify the idea of "natural" differences between the genders (Larrabee, 1993).

Overall, the body of evidence suggests that Kohlberg's stages themselves, and the instruments used to measure them, are not gender-biased (Snarey & Samuelson, 2014). At the same time, it has also been recognized that what Gilligan identified was an important additional component to morality, an ethic of care, that was not fully captured in Kohlberg's initial conceptualization of morality as based on principles of rights and justice (see Nunner-Winkler, 1993 for discussion of these two approaches to morality). This ethic of care is now generally accepted as a separate "ethical voice" (Snarey & Samuelson, 2014, p. 68) within morality, one that is used by both men and women, although contextual and gender differences in its prevalence or salience may exist. Mary Jeanne Larrabee (1993b), in her introduction to a volume exploring the ethic of care from feminist perspectives, notes that Gilligan herself was not seeking to replace Kohlberg's theory, but rather to add to it the types of focus on relational concerns that she heard from her female participants. Gilligan's work was pointing out that by developing his initial theory based on male samples, Kohlberg had inadvertently omitted the experiences of women, which were different from that of men, and which led to different, not necessarily deficit, emphases when considering moral dilemmas.

Yet of importance is also the methodological difference in Gilligan's work. Gilligan focused not on hypothetical scenarios, but on real life dilemmas. She studied moral reasoning as it existed in context. It is not unreasonable to think that this shift from the abstract to the concrete could change participants' understanding of and engagement in the task of moral reasoning. As I will discuss below, the assumption that a construct should remain stable across contexts is an assumption that we should question. And thus, the method itself must remain an aspect of the construct. In fact, Gilligan and her collaborators developed a specific method for analyzing interview data with a relational lens (Brown, Debold, Tappan, & Gilligan, 1991). Although initially developed as a tool for understanding women's lives, other researchers, in particular Niobe Way (2001), have used this method to understand the relational lives of boys and young men. I take from this history the lesson that researchers and practitioners alike need to be careful about how we define "normative" in relation to any construct, as all constructs are, at least in part, products of the worldviews of the researchers who develop them, the populations with which the researchers develop their measures, and the methods by which we collect and analyze our data.

Card's discussion of the development of the most widely used measures of gratitude reflect this issue of homogeneity of the samples on which instruments are developed. He rightly points out the concerns that come with a lack of diversity in populations studied when developing measures that may come to be widely used and, perhaps, even help define a field. In the case of gratitude, he notes, the most widely used measures have been tested in a variety of contexts, across diverse populations. This variety of contexts and populations increases researchers' confidence in the cross-context validity of the measures. Now we might argue that this is indeed the whole point of testing the validity and reliability of measures in different populations. But in measurement development, we tend to look for measures that perform *equivalently* across populations or that have been used with populations that are similar to those whom we are studying. Yet, as Card notes, psychometric properties are not properties of measures, but "are estimates of how a measure functions in a particular population, in a particular setting, and under particular methodological conditions" (Card, this volume, p. 33). Further, we don't always consider whether a given population's lower performance on a measure that has been demonstrated to be otherwise valid and reliable is a meaningful indicator of deficit or, instead, is an indicator that the construct itself is different in that context. This is particularly important because we all view the world from the bodies we inhabit. Kohlberg experienced the world as a White male, as did the subjects from whose data he first developed his theory. Gilligan experienced the world as a White female, as did the subjects from whose data she developed her theory. In both cases, there were still a lot people left out. While certainly non-White populations have been included in samples studying moral development since the theories were developed, their absence in the initial development means that what is defined as normative is largely still based on the experiences of White men and women, to say nothing of sample biases regarding social class, geography, sexuality, and other factors.

When researchers develop measures founded on definitions of constructs based in the experiences of a narrow subset of people, we risk defining the world based on the experiences and expectations of that subset. Which ideas, theories, or programs survive depends in part on who has the power to make their ideas heard and to get money to fund their ideas. At a convening of scholars discussing a *Futures of Children* volume about socioemotional development, a thought-provoking discussion occurred in response to the chapter by Anne Gregory and Edward Fergus. Their chapter focuses on racial disparities in school disciplinary practices and the links between these disparities and social emotional learning. In the chapter, Gregory and Fergus (2017) sug-

gest that a greater focus on teacher self-awareness is a necessary component of efforts to reduce the discipline gap between White and Black students. Further, they argue that "color blind" approaches to social emotional learning can inadvertently lead to understandings of socioemotional learning (SEL) competence that are themselves blind to diverse expressions of SEL from students from marginalized backgrounds. To create a more equity-oriented conception of SEL, the authors argue, the constructs underlying SEL, for both adults and children, should include factors such as an understanding of one's own conscious and unconscious biases as part of self-awareness or an awareness of the role of power and privilege in the actions of others as a part of social awareness. Consider how changing the definitions of those SEL constructs in line with these recommendations would lead to different measurements of those constructs, which may also lead to different populations being identified as in need of intervention.

Embedded in these last examples is the importance of context. Why do people have different experiences? Because context matters. As Card rightly points out, inter-informant reliability is only a reliable source of information if we assume that the construct itself is stable across contexts. Yet much human behavior is not. One study of African American adolescents found greater within-student variability in teacher perceptions of student behavior (cooperation, defiance, etc.) than between student variability. This could suggest either that the same student behaves differently with different teachers or that different teachers perceive the same behavior in different ways (Gregory & Thompson, 2010). In either case, a lack of inter-rater reliability between teachers may illustrate not instrument instability but actual differences in student and teacher experiences. Measurement equivalence assumes that a construct *should* operate "the same way across contexts, time, and/or groups" (Card, this volume, p. 32).

But if we believe that context matters, if we take Bronfenbrennner's (1979) social ecological theory seriously, then might we entertain the possibility that this is a false assumption? This is different than asking whether a person retains a core set of traits across contexts. It is asking if there is an essential center to the construct itself (see Gergen, 1985, for example). Or, if a core center exists, if it presents in different ways or means different things across contexts. Take selfishness, a value we may try to discourage in people, as an example. Kathryn Edin and Maria Kefalas's (2011) work with low-income mothers demonstrates how different people, with different life experiences, may define different behaviors as selfish. In their research, they found that some women from lower income backgrounds viewed upper income women's decisions to put off having children until they accomplished other life goals, including work and career goals, as "selfish and unnatural," (p. 208). Middle class narratives, however, typically paint the decision to have a child earlier in life, and outside of marriage, as selfish. So, whereas the *definition of selfish* for both sets of women may be the same (putting self before others), what is *identified as selfish behavior* differs based on social context and life experience. I recognize here that I am conflating the definition of a construct with the content of behaviors that may be identified with that construct. One may argue that outside of measures that rely on content-related items these context-dependent differences in content would not affect the measurement of character. But I contend that these differences matter because we tend to extrapolate from measures of specific constructs to judgments about individual-level (not context-specific) traits and what those traits represent about a person's underlying character. Further, we often take individuals' assessments of children, frequently based on categorizing of behaviors, at face value, without questioning the assumptions that underlie those assessments.

The interaction of construct definition and behavioral categorization, and the effects of

that interaction on youth, can be seen in issues related to school discipline, a topic discussed above. In a recent interview on the radio show *With Good Reason*, psychologist Jason Okonofua (2017) discussed his research on differential discipline practices by teachers for Black versus White students. Within the discussion, he makes an important point about the contextual specificity of how children's behavior is labeled. He provides an example of how, in the overcrowded, urban public schools he attended as a child, students' challenging or questioning teachers was viewed (i.e., defined or categorized) as insubordination and treated as such in terms of discipline. This same behavior, in the context of a small, private school he attended later, was encouraged, one may assume as a sign of some character-related trait, in this case, and in this context, a positive behavioral display. Whereas this is not a measurement issue per se, I would argue that the ways in which these differences play out in real world contexts are simply mirrors of how they play out in research.

Take, for example, the measurement of self-regulation. The famous marshmallow test has long been used as an indicator of self-regulation, a construct that is seen as part of social emotional development and that overlaps with elements of character (see Mischel, 2014). Over time, that indicator has proven to be highly predictive of other desirable outcomes, including academic achievement. Children who demonstrated poor performance on the marshmallow task were viewed as demonstrating less self-regulation, and therefore as being more impulsive, a character trait that was assumed to be in need of remediation. At the population level, children from lower on the socioeconomic ladder tended to perform "worse" on the marshmallow task. Various reasons were suggested for this, but the basic assumption was retained: that these youth, for some reason, were more impulsive and were less able to self-regulate. Yet recent research on children's physiological reactions during this task calls this assumption into question. Vagal tone is a physiological indicator of ability to self-regulate. It indicates how calm one is able to remain while under stress. Higher vagal tone should predict a stronger ability to delay gratification, i.e., to resist eating the marshmallow in order to hold out for the greater reward. In an experiment by Sturge-Apple and colleagues (2016), this is precisely what happened with children from middle class backgrounds. Yet in children from underresourced environments, the pattern did not hold. For those children, higher vagal tone was associated with decreased ability to delay gratification during the marshmallow task. Thus, for these children, the ability to remain calm while under pressure, one part of self-regulation, led to behavior that has been defined by researchers as an indication of poor self-regulation. Why this difference? The researchers speculate:

> Taken together, the present findings do not readily support conceptualizations drawn from normative psychological models proposing that high vagal tone is a marker of context-independent developmentally appropriate regulation. Instead, the current findings suggest that the function of high vagal tone may operate in a more curvilinear fashion toward facilitating behavioral fit within specific environmental contexts (Boyce & Ellis, 2005; West-Eberhard, 2003). In other words, high vagal tone may function differently within specific contexts, rather than having a context-independent function. (Sturge-Apple et al., p. 891).

In an interview with the *Washington Post* about this research (Ferdman, 2016), Sturge-Apple points out whereas we may think that we control for context when setting up a lab-based experiment, the outside environment still influences children's behaviors within those measurement settings. Importantly, I would add, when researchers come primarily from a particular set of backgrounds, we may be blind to the ways in which outside contexts can influence our measurements, thus falsely assuming a universality of experience of the measurement tool. Further, Sturge-Apple notes, the important question for researchers and practitioners who wish to intervene on

behalf of children becomes what the function of a behavior is in a particular context, not whether there is some universal quality to that behavior.

Of course, there are important reasons for social science to seek coherence of constructs across contexts. Complete relativism in a field would make generalized conclusions about human behavior and, therefore, scalable policy or programs, impossible. Yet we would be wise to recall Cronbach's (1975) cautions about generalizability in the social sciences:

> The aim of social and behavioral science, since Comte, has been to establish lawful relations comparable to those of the traditional natural sciences.… We need to reflect on what it means to establish empirical generalizations in a world in which most effects are interactive (Campbell, 1973). (Cronbach, 1975, p. 121)

Cronbach was responding to "the quest for nomothetic theory" (p. 116) that has dominated the field of psychology, and in particular the experimental testing of educational interventions. In the end, Cronbach (1975) concludes, "generalizations decay" (p. 122).

The quest for nomothetic theories reflects the fact that the application of knowledge by policymakers and programmers requires consistent definitions of constructs, tools to measure the same construct across contexts, and the ability to draw conclusions across contexts about the nature of humans in relation to a particular construct or intervention. These are very real needs. Yet real, too, is the prevalence of complex interactions that Cronbach highlighted as evidence that generalizations about relationships between constructs are tenuous. This produces a tension in social science, and here in the measurement of character as one dimension of human behavior, that must be acknowledged and addressed—not through dismissal of attempts to measure character, but instead through thoughtful interrogation of the stability and meaning of the constructs and relationships researchers do find. Diversification of the field in terms of *who* is conducting the research is one key to this endeavor. Researchers will be better able to identify where and when we may be misinterpreting a construct or an identified effect if we have individuals with diverse sets of background experiences to bring to bear on defining a construct or understanding a particular result.

Thus, I am not arguing that we should not try to quantitatively measure constructs such as character that we deem to be important to human development and to the social good. Rather, I am contending that we must recognize that any definition we create of a construct such as character, and any instrument we develop to measure it, reflects a shared understanding of that construct, not a reified, objective, and static reality to which we are holding a mirror. It is here that qualitative and mixed methods can complement quantitative measurements and help us identify and understand the implications of some of the potential measurement issues that arise when attempting to study character. The use of mixed methods can help both triangulate evidence, to identify and understand consistency across contexts, and examine where, how, and why, differences in constructs and relationships do exist.

Qualitative methods, alone or in combination with quantitative methods, help us understand the meanings that people make of their worlds. Qualitative methods are typically naturalistic and help us understand people's perceptions and lived experiences. Rather than focusing on measuring the quantity or prevalence of a construct or phenomenon, qualitative researchers focus on understanding its meaning as it exists in context (Guba & Lincoln, 1994). You can think about the strengths of qualitative social science as akin to qualitative analysis in chemistry. Chemical qualitative analysis is an analysis that aims to identify the components of a mixture. Likewise, qualitative social science is useful in understanding the components of a social setting, i.e., the context. So rather than having to omit outliers or accept some level of random error, qualitative research can help us understand the meaning of those outliers and perhaps the reasons for those variations (see Tolan & Deutsch,

2015 for discussion of the use of qualitative and mixed methods in developmental science).

Thus, qualitative methods allow us to dig into variation that may be masked, or identified but not explained, by quantitative-only work. For example, in a recent article on the consideration of race and ethnicity in research on youth development programs, Joanna Williams and I (Williams & Deutsch, 2016) argue that researchers need more nuanced attention to race and ethnicity in our work. In particular, our reliance on over-simplified racial and ethnic groupings, and on statistical comparisons between those groupings, likely masks important differences in individual experiences. For example, when we compare Black to Latinx to White youth on character development, we may miss important differences within each category. As we point out in our article, the category "Black" may include youth who are descendants of enslaved Americans, first generation immigrants from the Caribbean, or second generation African immigrants, just to name a few. Similarly, the category "Latinx" may lump together fourth generation Cuban youth with first generation Mexican youth with Honduran refugees. And "White" may include third generation Irish youth, second generation Jewish youth, and first generation Bosnian youth. Due to the requirements of statistical power and limited sample sizes, seldom can researchers interrogate differences within the categories we use to compare people. The overemphasis on between-group differences, we argue, ignores important within-group differences that help us understand how race and ethnicity operate as contexts of development that influence youth outcomes in different ways. Going back to my earlier point, qualitative methods could then help address issues of measurement equivalence by examining *how* and *why* measures may operate differently both within and between groups as well as over time, thus fine tuning both our understanding and measurement of a construct. In fact, whereas qualitative methods are often talked about in developmental work as exploratory, or as useful for measurement development, there is increasing acceptance that in-depth, rigorous qualitative and mixed methods work can yield important understandings about developmental pathways, or intraindividual change over time, and answer "why" and "how" questions by illuminating the ways in which a phenomenon occurs in the natural world (Tolan & Deutsch, 2015).

Anthropologists understand that we learn how to be part of a culture or society by observing and listening. Children learn by observing and listening, then analyzing, synthesizing, and applying the information they take in. So, if we want to learn how we become people of character, we, too, need to observe and listen, analyze and synthesize, which is what qualitative researchers do. Indeed, qualitative research in after-school settings, including that of Reed Larson, and others (e.g., Dawes & Larson, 2011; Larson & Brown, 2007; Wood, Larson, & Brown, 2009), has helped illuminate *how* youth develop skills, behaviors, and attitudes related to character in these settings. Observations of youth's relationships with adults and with each other have yielded important insights on what actions and structures promote character and other positive outcomes. And interviews with staff and youth have provided information on their perceptions, beliefs, and experiences. All of this information contributes to our understanding of how to structure and support organizations seeking to foster the character that most of us would agree is already *in* youth, but needs opportunities to flourish and be displayed.

And this is where I will leave you—not in a place of despair, existential or otherwise, but of hope. Hope that whereas character development is a broad, complex construct, we have a multitude of methods which we can use to understand it. When researchers and practitioners acknowledge that character development is a social construction and that context matters in its definition, exhibition, and measurement, when we apply critical consciousness[1] to our work, we can approach its study with more realistic expectations about what to expect from our research, and stronger tools to

help us gather the information we need to inform interventions and policy.

Acknowledgments: The author would like to thank the editors of this volume and the anonymous reviewers for their useful feedback on earlier versions of this article. I would also like to thank the National Academies of Science for inviting my participation in the meeting at which a version of this paper was first presented.

NOTE

1. "Critical consciousness," as described by Freire (1973), refers to the way in which individuals who are marginalized or oppressed within a society come to analyze the conditions in which they live and act to change those conditions. Youth development scholars have focused on critical consciousness as an important aspect of youth development, particularly for marginalized youth. Watts, Diemer, and Voight (2011), in their review of the field identify three components of critical consciousness: critical reflection, political efficacy, and critical action. Here, I purposefully invoke the critical consciousness of adults within the power structures of academia, youth programming, and policy. By remaining blind to the ways in which power structures shape social science research and its application, researchers and practitioners perpetuate, rather than promote change in, those systems of power. Thus, I argue, we should apply the tripartite components of critical reflection, political (or scholarly or vocational) efficacy, and critical action to our work.

REFERENCES

Anton, B. S. (2010). Proceedings of the American Psychological Association for the legislative year 2009: Minutes of the annual meeting of the Council of Representatives and minutes of the meetings of the Board of Directors. *American Psychologist, 65,* 385–475. doi:10.1037/a0019553

Baumrind, D. (1986). Sex differences in moral reasoning: Response to Walker's (1984) conclusion that there are none. *Child Development, 57,* 511-521. (Reprinted in *An ethic of care: Feminist and interdisciplinary perspectives,* pp. 177–192, by M. J. Larrabee, Ed., 1993, New York, NY: Routledge)

Berger, P., & Luckmann, T. (1966). *The social construction of knowledge: A treatise in the sociology of knowledge.* New York, NY: Open Road Media.

Berkowitz, M. W., & Bier, M. C. (2014). Research-based fundamentals of the effective promotion of character development in schools. In L. Nucci, T. Krettenauer, & D. Narvaez (Eds.). *Handbook of moral and character education* (2nd ed., pp. 248–260). New York, NY: Routledge.

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

Brown, L. M., Debold, E., Tappan, M., & Gilligan, C. (1991). Reading narratives of conflict and choice for self and moral voices: A relational method. In W. M. Kurtines & J. L. Gerwitz (Eds.). *Handbook of moral behavior and development: Vol. 2. Research* (pp. 25-61). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 12*(11), 116–127.

Dawes, N. P., & Larson, R. (2011). How youth get engaged: Grounded-theory research on motivational development in organized youth programs. *Developmental Psychology, 47*(1), 259–269.

Edin, K., & Kefalas, M. (2011). *Promises I can keep: Why poor women put motherhood before marriage.* Los Angeles, CA: University of California Press.

Ferdman, R. A. (2016, June 8). The big problem with one of the most popular assumptions about the poor. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/wonk/wp/2016/06/08/the-problem-with-one-of-the-most-popular-assumptions-about-the-poor/?utm_term=.eea45448ebca

Freire, P. (1973). *Education for critical consciousness.* New York, NY: Seabury.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist, 40*(3), 266–275.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Gregory, A. & Fergus, E. (2017). Social and emotional learning and equity in school discipline. *The Future of Children, 27*(1), 117–136.

Gregory, A., & Thompson, A. R. (2010). African American high school students and variability in behavior across classrooms. *Journal of Community Psychology, 38*(3), 386–402.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: SAGE.

Karasic, D., & Drescher, J. (2006). Sexual and gender diagnoses of the diagnostic and statistical manual (DSM): A reevaluation. *Journal of Psychology & Human Sexuality, 17*(3-4), 1–5.

Kohlberg, L. (1976). Moral stages and moralization. In T. Lickona (Ed.), *Moral development and Behavior* (pp. 31–53). New York, NY: Holt, Rinehart & Winston.

Kohlberg, L. (1981). *Essays on moral development: Vol. 2. The psychology of moral development*. San Francisco, CA: Harper & Row.

Larrabee, M. J. (Ed.). (1993). *An ethic of care: Feminist and interdisciplinary perspectives.* New York, NY: Routledge.

Larson, R. W., & Brown, J. R. (2007). Emotional development in adolescence: What can be learned from a high school theater program? *Child Development, 78*(4), 1083–1099.

Latour, B. (1991). *We have never been modern.* Cambridge, MA: Harvard University Press.

Mischel, W. (2014). *The marshmallow test: Why self-control is the engine of success.* New York, NY: Little, Brown & Company.

Nunner-Winkler, G. (1993). Two moralities? A critical discussion of an ethic of care and responsibility versus an ethic of rights and justice. In M. J. Larrabee (Ed.), *An ethic of care: Feminist and interdisciplinary perspectives* (pp. 143–156). New York, NY: Routledge.

Okonofua, J. (2017, May 26). Discipline disparities. [Interview]. *With Good Reason.* A program of the Virginia Foundation for the Humanities. Retrieved from http://withgoodreasonradio.org/episode/degrees-of-separation-secondary/?t=00:28:01

Seider, S., Jayawickreme, E., & Lerner, R. M. (2017). Theoretical and empirical bases of character development in adolescence: A view of the Issues. *Journal of Youth and Adolescence, 46*, 1149–1152. doi:10.1007/s10964-017-0650-3

Snarey, J., & Samuelson, P. L. (2014). Lawrence Kohlberg's revolutionary ideas: Moral education in the cognitive-developmental tradition. In L. Nucci, D. Narveaz, & T. Krattenauer (Eds.) *Handbook of moral and character education* (pp. 61–83). New York, NY: Routledge.

Sturge-Apple, M. L., Suor, J. H., Davies, P. T., Cicchetti, D., Skibo, M. A., & Rogosch, F. A (2016). Vagal tone and children's delay of gratification: Differential sensitivity in resource-poor and resource-rich environments. *Psychological Science, 27*(6), 885–893.

Tolan, P. H., & Deutsch, N. L. (2015). Mixed methods in developmental science. In W. F. Overton & P. C. Molenaar (Eds.), *Handbook of child psychology and developmental science: Vol. 1. Theory and method* (7th ed., pp. 713–757). Hoboken, NJ: Wiley.

Walker, L. J. (1984). Sex differences in the development of moral reasoning: A critical review. *Child Development, 55,* 677–691. Reprinted in M.J. Larrabee (Ed.) *An Ethic of Care: Feminist and Interdisciplinary Perspectives.* (157–176). New York: Routledge. (Reprinted in *An ethic of care: Feminist and interdisciplinary perspectives,* pp. 157–176, by M. J. Larrabee, Ed., 1993, New York, NY: Routledge)

Watts, R. G., Diemer, M. A., & Voight, A. M. (2011). Critical consciousness: Current status and future directions. In C. A. Flanagan & B. D. Christens (Eds.), *Youth civic development: Work at the cutting edge. New Directions for Child and Adolescent Development, 134,* 43–57.

Way, N. (2001). Using feminist research methods to explore boys' relationships. In D. L. Tolman & M. Brydon-Miller (Eds.), *From subjects to subjectives: A handbook of interpretive and participatory methods* (pp. 111–129). New York, NY: New York University Press.

Williams, J. L., & Deutsch, N. L. (2016). Beyond between-group differences: Considering race, ethnicity, and culture in research on positive youth development programs. *Applied Developmental Science, 20*(3), 203–213. doi:10.1080/10888691.2015.1113880

Wood, D., Larson, R.W., & Brown, J. R. (2009). How adolescents come to see themselves as more responsible through participation in youth programs. *Child Development, 80*(1), 295–309.

Yoshikawa, H., Weisner, T. S., Kalil, A., & Way, N. (2008). Mixing qualitative and quantitative

research in developmental science: Use and methodological choices. *Developmental Psychology, 44*, 344–354. doi:10.1037/0012–1649.44.2.344