Philosophy of Science (2024), **91**, 1467–1477
doi:10.1017/psa.2023.147

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Path-Dependence in Measurement: A Problem for Coherentism

Morgan Thompson

Department of Philosophy, Cornell University, Ithaca, NY, USA
Email: morganthompson28@gmail.com

(Received 11 April 2023; revised 25 September 2023; accepted 05 October 2023; first published online 20 October 2023)

Abstract

Racial discrimination is a multidimensional concept. Yet, subjective measures of racial discrimination focus on particular dimensions (interpersonal over institutional, attributable over ambiguously attributable). I argue that there is path dependence in the development and validation of subjective measures, such that existing scales prevent the development of scales that are better for some purpose. Path dependence can occur when researchers: (1) adopt a coherentist view of measurement, namely, in iteratively refining constructs and measures and (2) employ current psychometric validation practices. The main takeaway is that norms are needed to evaluate the initial refinement of the construct rather than taking it for granted.

1. Introduction

Social sciences investigate multifaceted phenomena that are characterized by paradigmatic examples or by qualitative dimensions rather than definitions. However, during measurement researchers must reduce their dimensionality. As a result, the construct is revised and precisified so that attention is focused on particular dimensions and away from others. Sometimes whole research traditions adopt the same ways of revising the construct in pursuit of its measurement, leaving important properties or dimensions ignored or abandoned. Call it path dependence when it is more suitable for some purposes to measure these other dimensions of the phenomenon, but the prevalence of the current measures disadvantages the development of new measures.

Path dependence describes the situation for measuring racially discriminatory experiences in fields like public health, psychology, and sociology. Measures of racial discrimination focus on particular dimensions of the phenomenon (interpersonal over institutional features, clear over ambiguous attributability), despite the recognition that our folk and theoretical understandings of racial discrimination are multidimensional.

I argue that path dependence can occur when researchers: (1) adopt coherentist views of measurement, namely, iteratively refining constructs and their measures

and (2) employ current validation practices in psychometrics. The main takeaway is that researchers need to evaluate the initial formation of the construct rather than taking it for granted. That is, coherentist approaches ought to include norms directing researchers to reconsider the dimension(s) of the construct that were the initial focus, not merely norms about self-correcting the relationship between the measurement scale and the revised construct.

2. Measures of racially discriminatory experiences

The theoretical concept of racial discrimination is composed of a number of cross-cutting dimensions: interpersonal, internalized, institutional, direct, clearly attributable, subtle, and so forth. Broadly, racial discrimination is unfair treatment or impact on the basis of race. Interpersonal racial discrimination occurs when an interaction between individuals involves unfair treatment or impact. Racial discrimination can also occur institutionally when policies, laws, and institutions produce unfair outcomes. In some cases, interpersonal racial discrimination may also be institutional discrimination, such as when a police officer qua officer engages in unfair treatment on the basis of race.

Let us examine self-report measures of racial discrimination (see Table 1).

According to researchers, these scales are reliable and well validated (e.g., Bastos et al. 2010). Yet, they focus on cases in which the person reporting is able to clearly attribute the experience as caused by their membership in some racial group and thus, do not measure ambiguous cases of discrimination (Thompson 2023). Further, scales tend to focus on interpersonal discrimination over institutional discrimination. When an item concerns institutional features (such as being denied a loan), the item is either mixed with interpersonal items (as in Krieger's scale) or in wholly distinct scales (Williams's Major Discrimination Scale). Consider the latter scale that includes questions about being unfairly stopped by the police or being unfairly fired. The scale aims to measure these severe (or "major") discriminatory events that would disrupt individual's lives. Mundane institutional discriminatory events are not measured (e.g., lacking access to preventative healthcare).

However, these scales all make the *same* choices about which dimensions to include and which to neglect. This raises the question: Why are certain dimensions, such as ambiguous attribution and institutionality, largely ignored in subjective measures?

3. Path dependence

Path dependence describes the development of some technology that promotes particular theoretical and design choices at the expense of others. Path dependent systems are those in which some change in a system's history reinforces its development along particular paths and hinders others. The concept of path dependence gained traction in science and technology studies (e.g., David 2001) to analyze why some standards set by a previous technology are perpetuated in future designs, even when alternatives are equivalent or better. A classic but contested example is the QWERTY keyboard layout (David 1985; Liebowitz and Margolis 1990). According to David (1985), the QWERTY layout was chosen to solve a technical issue with previous typewriter designs, namely, putting commonly used letters further apart to avoid errors when congruent type bars are pressed in sequence. While we no

Table 1. Example subjective measures of experiences of racial discrimination

Krieger et al. (2005) Experiences of Discrimination Scale	Have you ever experienced discrimination, been prevented from doing something, or been hassled or made to feel inferior in any of the following situations because of your race, ethnicity, or color? (1) At school? (2) Getting hired or getting a job? (3) At work? (4) Getting housing? (5) Getting medical care? (6) Getting service in a store restaurant? (7) Getting credit, bank loans, or a mortgage? (8) On the street or in a public setting? (9) From the police or in the courts?
Landrine et al. (2006) Schedule of Racist Events (SRE)	Example Questions: "How many times have you been treated unfairly by [your coworkers, fellow students, and colleagues; people in service jobs; strangers; people in helping jobs; neighbors; institutions] because you are Black?" How many times in the past year? "Never" to "almost all the time" How many times in your entire life? "Never" to "almost all the time" How stressful was this for you? "Not at all" to "extremely"
Williams et al. (1997) Everyday Discrimination Scale	In your day-to-day life, how often do any of the following things happen to you? 1. You are treated with less courtesy than other people are. 2. You are treated with less respect than other people are. 3. You receive poorer service than other people at restaurants or stores. 4. People act as if they think you are not smart. 5. People act as if they are afraid of you. 6. People act as if they think you are dishonest. 7. People act as if they're better than you are. 8. You are called names or insulted. 9. You are threatened or harassed. What do you think is the main reason for these experiences? 1. Your ancestry or national origins 2. Your gender 3. Your race 4. Your age 5. Your religion 6. Your height 7. Your weight 8. Some other aspect of your physical appearance 9. Your sexual orientation 10. Your education or income level

longer face this technical problem, QWERTY layouts persist even when superior alternatives (e.g., Dvorak) exist.

I propose that a research tradition exhibits path dependence to the extent that the following conditions are met: (1) existing technologies are deficient *for some purpose*, (2) alternative technologies could be proposed that are superior *with respect to that purpose*, but (3) the alternatives are disadvantaged in use relative to the existing technologies (*e.g., are ignored, receive less uptake in future research, are used by a small*

population of potential users). Part of these criteria are descriptive in that it must be demonstrated that the range of existing technologies make similar assumptions about the theoretical construct (that describes what is measured) and that there are disadvantages to alternative technologies. These criteria are also normative because it must be shown that existing technologies are deficient for some purpose, while alternative technologies would be better.

3.1. Deficiency

Here I apply the first two criteria to the case. The first criterion holds that existing technologies are deficient for some purpose. In the case of discrimination scales, the deficiency is in part based on common assumptions about the relevant dimensions of the theoretical construct. Existing scales do not measure certain kinds of racially discriminatory experiences, namely, those that are institutional but noninterpersonal factors (such as policies) and those that cannot be clearly attributed to race. Reducing dimensionality of complex social constructs is often a necessary part of social science. It does not by itself demonstrate deficiency. It must be shown that, for some particular purpose, it is deficient to merely employ the existing measures. In other words, do we have purposes for which it would be useful to measure other dimensions?

Here I want to claim that existing scales are deficient for two purposes: (1) understanding the causes of social alienation and (2) the integration of evidence about interpersonal and institutional discrimination. I address each in turn.

Some experiences of discrimination lack clear actors or cannot be clearly attributable to one's social identity; these experiences may be more associated with feelings of social alienation than cases that are clearly caused by, for example, racial hatred. Here are some suggestive reasons in support of this claim. Researchers have found that high self-identification with one's social group can prevent some negative effects of racial discrimination (e.g., Mossakowski 2003; Sellers et al. 2003). This evidence is consistent with high self-identification playing a role in the appraisal of the experience. Perhaps individuals with high self-identification with a particular racial group are more likely to interpret cases of discrimination as clearly attributable to their race, but also to dismiss the negative evaluations of their racial group that accompany this treatment. Individuals with low self-identification may be less likely to report discrimination because their experiences of discrimination will be less clearly attributable to race. This could be one cause of social alienation from social groups in general. Measures of ambiguously attributable discrimination are needed to test the potential mediation of the relationship between self-identification with a racial group and social alienation.

The second purpose for subjective measures of discrimination could be to integrate evidence concerning interpersonal and institutional discrimination. Currently there are different methods for measuring discrimination that broadly track either interpersonal discrimination or institutional discrimination. Subjective measures tend to measure interpersonal interactions and their effects (excluding scales of Major Discrimination). Institutional discrimination is often measured using large-scale population studies or field experiments. Large-scale population studies assess

the impact of policy changes or other environmental factors on different racial groups. Field experiments investigate when individuals are treated differently due to race by examining response rates to nearly identical individuals (except for race) in various decision procedures (e.g., housing, employment). However, some institutional discrimination is mediated by interpersonal interactions. Some agent acts as a representative and enforcer of institutional policies, such as police officers who escalate interactions with the public on the basis of probable cause that disproportionately impacts Black and Hispanic populations. Subjective measures concerning institutional-interpersonal discrimination could help researchers integrate evidence about these dimensions of discrimination. While Williams's Major Discrimination scale does this on some items, other items are left open to interpretation (e.g., denied a bank loan by whom?). More intentional design of the scale to identify experiences with both interpersonal and institutional dimensions would better serve integrative goals.

3.2. Superior alternatives

The second criterion for path dependence is that alternative technologies are superior (compared to existing technologies) with respect to some particular goal. Here we can again appeal to the goals of explaining social alienation and integrating evidence about institutional discrimination that occurs interpersonally. Alternative subjective measures of ambiguous discrimination and of institutional discrimination that occurs using representatives of the institution would better serve these goals than existing measures. Creating measures of ambiguous discrimination will likely need a different approach than directly asking individuals to report experiences of discrimination, which will favor reports of clear cases. Instead, a subjective measure may offer vignettes of putative cases of ambiguous discrimination and measure interpretations of the case along with other relevant variables (i.e., self-identification with group). However, these alternative subjective measures, to my knowledge, do not yet exist. Thus, the second criterion is only partially met. Were such alternative measures to exist, then they would better serve some research purposes.

3.3. Disadvantaged alternatives

The third criterion for path dependence requires that the alternative technologies are disadvantaged relative to existing technologies. It is this criterion that demonstrates how path dependence is problematic for research. Despite having particular goals for which alternative measures would be better, these measures are not developed or ignored. For example, disadvantage could be due to institutional factors (such as the prevalence of training programs for employees using QWERTY layouts), social factors (such as devaluing the tools created based on testimony from marginalized groups; Wu 2023), or factors concerning philosophy of measurement. In the next two sections, I argue that adopting a coherentist view of measurement (section 4) along with current validation practices in psychometrics (section 5) can produce this disadvantage. I conclude that the contemporary coherentists approach is incomplete, not that it is incorrect.

4. Coherentist picture of measurement

Early views of measurement were foundationalist in that the development and validation of measurement tools were taken to be primarily empirical matters (such as Bacon's view on which theories can be constructed from generalizing on the basis of unassailable facts). Foundationalism assumes that researchers have independent access to the material being measured. When evaluating a new measure, the instrument provides marks that are approximations of the "true values." Justification of claims is given by empirical facts about the object of interest. Through theorizing on the foundations of these facts, gradually an understanding of both the measurement tools and construct could be achieved. However, the claim that researchers have access to self-justifying claims about the "true values" of the quantities they measure has largely failed (Chang 2007; Tal 2017b).

Let us distinguish between instrument indications, which are the material effects on the measurement instrument after the measurement process is finished, and measurement outcomes, which are knowledge claims based on the indications and background knowledge about the measured phenomenon (Tal 2017a). Coherentists have argued that neither instrument indications nor measurement outcomes provide self-justifying claims that can be foundational. Instrument indications are produced by a combination of the environment, phenomenon, and the instrument; they provide evidence for the measurement outcome claims but are not themselves given. For example, a faulty instrument may produce instrument indications that do not track the phenomenon.

Coherentism, however, accepts that both empirical and theoretical considerations play a role in measurement. Initial theoretical assumptions can be tentatively assumed. Then these assumptions allow researchers to begin an iterative investigation whereby both the measurement process and the theoretical assumptions are revised to better align. Unlike foundationalism, the initial theoretical assumptions are contingent starting points that are subject to revision.

Coherentists claim to provide two benefits over foundationalist views. First, coherentists views better describe scientific practice, including both historical episodes (Chang 2004) and contemporary practice (Tal 2017a). Second, coherentists demonstrate how the justification of measurement outcomes, namely, that these knowledge claims are reliable and valid, can be possible in the absence of appeal to "true values." Instead, sufficiently reliable and valid knowledge claims can be justified by appealing to a process of epistemic iteration.

Justification is achieved by building a body of knowledge on a tentative foundation that is refined and self-corrected. The major normative force of progressive coherentism is epistemic iteration (Chang 2004, 2007). Chang discusses two methods of epistemic iteration: conceptual enrichment and self-correction. Enrichment of the concept involves making the concept more refined, precise, and/or extending it to new domains. For example, Chang (2004) traces the enrichment of the temperature concept through its quantification. Temperature was initially only a qualitative concept that came to be measured by thermoscopes. As a result, comparisons between instances could be achieved and an order was imposed on thermal sensations. Numerical thermometers were based on this ordering of sensations. Iterative

refinement of thermoscopes attached a cardinal scale to the ordering that allowed measurement in degrees.

Epistemic iteration also involves self-correction. Suppose some standard is accepted for initial measurement. Later standards developed by epistemic iteration can be used to correct and replace the first standard. Chang (2007) illustrates with the example of imperfect glasses. A pair of glasses may be scratched or distorted so that it produced distorted vision, but these errors may not be seen until we first use the glasses to observe them (i.e., in a mirror). First we trust the glasses provisionally because we (or some of us) see the visual scene more clearly with them. But we also observe the defects in the visual scene observed through them. Once recognized, we can correct for these errors. Thus, an initial standard (to provisionally trust the precise visual scene presented by the glasses) is overruled by a later standard (to trust the precise visual scene except where we have noted distorting features of the glasses), but this replacement relies on the temporary adoption of the initial standard in the absence of foundationalist warrant.

5. Evaluating the validity of new measures

One plausible solution to the problem is to diversify the range of available measures. New measures could examine different dimensions of racial discrimination and, thus, treat different claims about the construct as fallible standards. I argue here that it is difficult to produce diverse alternative measures due to current validation practices.

Validation assesses the relationship between the test and the construct. Researchers can evaluate the extent to which the questionnaire items fit with the dimensions of the construct based on our theoretical understanding. Content validity is the extent to which the measurement procedure reflects all and only the dimensions of the construct. The relevant dimensions of the construct come from other sources of evidence, such as theoretical conceptions of it (Krieger 2014) or dimensions of the folk concept (Bloch-Mullins 2022; Feest 2017). It is evaluated conceptually with respect to the content of the survey items.

Researchers can also evaluate the extent to which questionnaire items seem to produce indications (or scores) of the same phenomenon. Construct validity is the extent to which scales measure the intended phenomenon and not unrelated phenomena (Campbell and Fiske 1959). One aspect of construct validity is convergent validity, or the extent to which scales intended to measure the same construct do so. It is evaluated empirically with respect to the relations among responses to sets of survey items.

5.1. Content validity

Bastos et al. (2010) is the only meta-analysis of validation practices concerning subjective racial discrimination scales. They identified twenty-four scales, twenty-one of which assessed content validity and/or convergent validity.

Bastos and colleagues' (2010) found that the majority of the items on the scales were developed through literature review (15/24 studies). That is, on the basis of theoretical and empirical literature on subjective experiences of racial discrimination, the relevant dimensions to be measured were identified. Fewer studies developed items based on the population's understanding of the construct or through

empirical investigation (focus group: 6 of 24; interview: 5 of 24; empirical investigation: 5 of 24), which would allow for dimensions not already published in the literature to be included. Once the scale was developed and used with a study population, its content validity can be evaluated. According to Bastos et al. (2010), only six of the twenty-four scales evaluated content validity using focus groups, pretests, pilot studies, or interviews. These evaluations are more open to including alternative dimensions of the construct. Four of twenty-four scales assessed content validity using literature reviews or expert panels, which is much more likely to evaluate scales relative to the dimensions of the revised construct.

Together the development of items and assessment of content validity means that decisions about which construct dimensions to measure in the first place are propagated in later scales. An early scale is developed using a construct that emphasizes certain dimensions and neglects others. This common for social science concepts that lack precise definitions (Bradburn et al. 2017), but it is also a feature of the coherentist picture that some dimensions of constructs get removed as the construct is refined. By relying on previous research to develop items, this privileges the dimensions chosen early over others regardless of the strength of reasons to focus on those dimensions. In combination with assessing content validity using the dimensions of the revised construct, there is little opportunity to include new dimensions of the construct.

There are a few reasons why some revised construct may be too narrow. First, the revised construct and the scale may ignore dimensions that are relevant for some purposes. Second, the revised construct may differ too significantly from the original theoretical construct, especially in cases in which the construct we seek to measure involves subjective experiences. However, if content validity is not assessed (through interviews, focus groups, and testimony by participants), then researchers may not recognize when the revised construct is too narrow (Cupples 2017; McClimans 2017) or when there is a mismatch between participants' and researchers' understandings of a construct (Cupples 2017, 109).

5.2. Convergent validity

Convergent validity is often assessed by examining correlations among the indications of the scales. Sixteen of the twenty-four scales assessed convergent validity, which means they assessed the extent to which scores on the new scale were positively correlated with scores on existing scales (Bastos et al. 2010).

This indicates that researchers take these scales to measure the same phenomenon. By itself, this does not amount to path dependence. However, it may contribute to the extent that construct validity is taken to be an important source of information about scale quality. If high convergent validity is needed for the community to agree that a new scale is valid, then researchers will seek new scales that produce scores similar to existing measures. Thus, this provides another avenue for path dependence.

5.3. Same construct?

Here I have treated discrimination as a single construct with multiple dimensions. This is in line with researchers' conceptions of discrimination as being measured in

different ways (e.g., subjective measures, statistical analyses, audit studies). Could it also be that these different kinds of measures get at distinct constructs? First, we should avoid reifying the measurement process as features of the construct. This principle was important in the debates about indirect measures and implicit attitudes. Subjective measures might be used to measure interpersonal dimensions or institutional dimensions of discrimination. Second, we might ask whether interpersonal and institutional dimensions warrant treatment as different constructs. Some hold that dissociative evidence (concerning some construct dimensions) warrants that a construct be split into distinct constructs (Feest 2011). However, there are some reasons to be cautious of this response. First, there is not (yet) evidence of dissociation between the dimensions because there is not (yet) a diversity of measures. Second, dissociation among dimensions can sometimes be useful in construct development that later results in a new unified construct (Bloch-Mullins 2022). Proponents of this objection should still accept my arguments that a diversity of measures is needed.

6. Conclusion

I argued that path dependence in subjective measures constitutes a problem. However, there are other measures of discrimination, such as statistical analyses of observational data. Perhaps ambiguous and institutional discrimination are sufficiently captured by these measures and thus, new subjective measures are not needed.

Note that subjective measures and statistical analyses are subject to different biases. Subjective measures may have reporting biases, whereas statistical analyses (e.g., economic studies of wealth distribution) may underestimate racial discrimination. Statistical studies rely on the assumption that when all other variables are controlled for (e.g., socioeconomic status), then any difference by race in some outcome variable is due to racial discrimination. Lily Hu (2023) makes a convincing argument (in the context of predictive algorithms) that the variables controlled for (e.g., arrest rates) partly constitute the social constructivist concept of race (rather than merely correlating with race). If socioeconomic status partly constitutes race, then economic studies of discrimination alone would tend to underestimate racial discrimination. Subjective measures do not control for these variables and are not subject to these biases.

Still, it does not mean that subjective measures are better than statistical analyses for the purpose of estimating racial discrimination because that depends on the extent of reporting biases and amount of unperceived discrimination. Rather, it is often better to have multiple measures of the same construct (or construct dimension). Where both measures investigate institutional discrimination, their results can be compared (i.e., triangulation). When the results agree, we can have increased confidence in them; when they disagree, we investigate which biases skew the results. Rather than assuming statistical analyses provide better estimates of discrimination, we should diversify the measures available to (dis)confirm our hypotheses.

One might also think that pluralism can offer an antidote to the problem of path dependence. As Chang (2017, 233) notes, "epistemic iteration allows the flourishing of

competing traditions, each of which can progress on its own basis without always needing to be judged in relation to others." This quote clarifies two things: First, coherentist epistemic iteration would avoid the problems of current validation practices. Second, and more relevant here, on the coherentist picture researchers are *permitted* to choose different starting points. Pluralism arises when different researchers choose different ways of developing a construct from the multifaceted phenomenon. However, on the coherentist picture, there is no *requirement* or *incentive* to set out in new directions and produce pluralism.

Coherentist approaches to measurement need norms for evaluating the process of transforming a construct into a revised construct. Without them, the coherentist approach is incomplete. It cannot evaluate research projects that leave a productive path untaken. Which qualitative dimensions are worthy of new measures? When should some characteristic of an everyday concept be abandoned in measurement? The current coherentist picture leaves us without answers to these questions.

References

Bastos, Joao L., Celeste, Roger K., Faerstein, Eduardo, and Barros, Aluisio J. D. 2010. "Racial Discrimination and Health: A Systematic Review of Scales with a Focus on Their Psychometric Properties." *Social Science & Medicine* 70 (7):1091–99. https://doi.org/10.1016/j.socscimed.2009.12.020

Bloch-Mullins, Corinne L. 2022. "Similarity in the Making: How Folk Psychological Concepts Facilitate Development of Psychological Concepts." *Synthese* 200 (2):76. https://doi.org/10.1007/s11229-022-03600-6

Bradburn, Norman M., Cartwright, Nancy, and Fuller, Jonathan. 2017. "A Theory of Measurement." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, 73–88.

Campbell, Donald, and Fiske, Donald. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (2):81–105.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chang, Hasok. 2007. "Scientific Progress: Beyond Foundationalism and Coherentism." *Royal Institute of Philosophy Supplements* 61:1–20.

Chang, Hasok. 2017. "VI—Operational Coherence as the Source of Truth." *Proceedings of the Aristotelian Society* 117 (2):103–22. https://doi.org/10.1093/arisoc/aox004

Cupples, Laura. 2017. "The Epistemological Roles of Models in Health Science Measurement." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, edited by Leah McClimans, 107–20. Rowman & Littlefield Publishers.

David, Paul A. 2001. "Path Dependence, Its Critics, and the Quest for 'Historical Economics.'" In *Evolution and Path Dependence in Economic Ideas: Past and New*, edited by P. Garroutse and S. Ioannides, 120–142. Edward Elgar Publishing.

David, Paul A. 1985. "Clio and the Economics of QWERTY." *The American Economic Review*, 75 (2):332–37.

Feest, Uljana. 2011. "Remembering (Short-Term) Memory: Oscillations of an Epistemic Thing." *Erkenntnis* 75 (3):391–411. https://doi.org/10.1007/s10670-011-9341-8

Feest, Uljana. 2017. "Phenomena and Objects of Research in the Cognitive and Behavioral Sciences." *Philosophy of Science* 84 (5):1165–76. https://doi.org/10.1086/694155

Hu, Lily. 2023. "What Is 'Race' in Algorithmic Discrimination on the Basis of Race?" *Journal of Moral Philosophy* 1–26. https://doi.org/10.1163/17455243-20234369

Krieger, Nancy. 2014. "Discrimination and health inequities." *International Journal of Health Services* 44 (4):643–710. https://doi.org/10.2190/HS.44.4.b

Krieger, Nancy, Smith, Kevin, Naishadham, Deepa, Hartman, Cathy, and Barbeau, Elizabeth M. 2005. "Experiences of Discrimination: Validity and Reliability of a Self-Report Measure for Population Health Research on Racism and Health." *Social Science & Medicine* 61 (7):1576–96. https://doi.org/10.1016/j.socscimed.2005.03.006

Landrine, Hope, Klonoff, Elizabeth A., Corral, Irma, Fernandez, Senaida, & Roesch, Scott. 2006. "Conceptualizing and Measuring Ethnic Discrimination in Health Research." *Journal of Behavioral Medicine* 29 (1):79–94. https://doi.org/10.1007/s10865-005-9029-0

Liebowitz, Stan J., and Margolis, Stephen E. 1990. "The Fable of the Keys." *The Journal of Law and Economics* 33 (1):1–25. https://doi.org/10.1086/467198

McClimans, Leah. 2017. "Psychological Measures, Risk, and Values." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, edited by Leah McClimans. Rowman & Littlefield Publishers. 89-106.

Mossakowski, Krysia N. 2003. "Coping with Perceived Discrimination: Does Ethnic Identity Protect Mental Health?" *Journal of Health and Social Behavior* 44 (3):318-331. https://doi.org/10.2307/1519782

Sellers, Robert M., Caldwell, Cleopatra H., Schmeelk-Cone, Karen H., and Zimmerman, Marc A. 2003. "Racial Identity, Racial Discrimination, Perceived Stress, and Psychological Distress among African American Young Adults." *Journal of Health and Social Behavior* 44 (3):302-317. https://doi.org/10.2307/1519781

Tal, Eran. 2017a. "Calibration: Modelling the Measurement Process." *Studies in History and Philosophy of Science Part A* 65–66:33–45. https://doi.org/10.1016/j.shpsa.2017.09.001

Tal, Eran. 2017b. "Model-Based Epistemology of Measurement." In *Reasoning in Measurement*, 245–65. Routledge.

Thompson, Morgan. 2023. "Characterizing and Measuring Racial Discrimination in Public Health Research." *Philosophy of Science* 1–32. https://doi.org/10.1017/psa.2023.51

Williams, David R., Yu, Yan, Jackson, James S., and Anderson, Norman B. 1997. "Racial Differences in Physical and Mental Health: Socio-economic Status, Stress and Discrimination." *Journal of Health Psychology* 2 (3):335–51. https://doi.org/10.1177/135910539700200305

Wu, Jing-yi. 2023. "Epistemic Advantage on the Margin: A Network Standpoint Epistemology." *Philosophy and Phenomenological Research* 106 (3):755–777. https://doi.org/10.1111/phpr.12895

Cite this article: Thompson, Morgan. 2024. "Path-Dependence in Measurement: A Problem for Coherentism." *Philosophy of Science* 91 (5):1467–1477. https://doi.org/10.1017/psa.2023.147