

## Supervisees' of differing genders and races perceptions of power in supervision

Stefanie A. Wind, Ryan M. Cook & W. Bradley McKibben

**To cite this article:** Stefanie A. Wind, Ryan M. Cook & W. Bradley McKibben (2021) Supervisees' of differing genders and races perceptions of power in supervision, *Counselling Psychology Quarterly*, 34:2, 275-297, DOI: [10.1080/09515070.2020.1731791](https://doi.org/10.1080/09515070.2020.1731791)

**To link to this article:** <https://doi.org/10.1080/09515070.2020.1731791>



Published online: 20 Feb 2020.



Submit your article to this journal [↗](#)



Article views: 675



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

ARTICLE



## Supervisees' of differing genders and races perceptions of power in supervision

Stefanie A. Wind <sup>a</sup>, Ryan M. Cook <sup>a</sup> and W. Bradley McKibben<sup>b</sup>

<sup>a</sup>Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, Tuscaloosa, USA; <sup>b</sup>Department of Counseling, Nova Southeastern University, Fort Lauderdale, USA

### ABSTRACT

In the current study, we examined the extent to which supervisees' perceptions of power dynamics related to gender and race in a sample of 229 trainees. Overall, we did not find systematic differences in supervisees' perceptions of power in clinical supervision based on their gender and race. However, utilizing differential item functioning (DIF) analyses, we found evidence that female and male supervisees perceived power differently for specific aspects of power in clinical supervision. Female supervisees perceived their supervisors as possessing more power in identifying goals of clinical supervision, conceptualizing client cases, and initiating discussions of the power dynamics in the supervisory relationship. Male supervisees perceived their supervisors as possessing more power in providing feedback about their clinical work and counseling skills. Regarding race, we found only slight-to-moderate DIF for one item, Item 10 (i.e. feedback on work with clients"). In light of small sample sizes for some groups, we also examined model-data fit for individual supervisees. These analyses allowed us to explore the degree to which individual supervisees interpreted power dynamics consistently with the larger sample. We identified individual supervisees for whom model estimates had different interpretations from the larger sample. Implications for supervisors and supervision scholars are discussed.

### ARTICLE HISTORY

Received 12 October 2019  
Accepted 15 February 2020

### KEYWORDS

Clinical supervision; power dynamics; differential item functioning; person fit; rating scale

Due to the evaluative nature of clinical supervision, power dynamics between supervisors and supervisees are inescapable. In the context of clinical supervision, *power dynamics* reflect the distribution of power between supervisors and supervisees, and such dynamics likely change and shift throughout the dyad's working relationship, with supervisees or supervisors sometimes holding more or less power for specific aspects of supervision or sharing the power for other aspects. Supervisors are empowered in their roles as evaluators and gatekeepers, while supervisees are called openly to discuss personal and clinical concerns and to receive their supervisors' feedback (Bernard & Goodyear, 2019). Power in clinical supervision is not fundamentally problematic (Cook et.al, 2018 ; Murphy & Wright, 2005), but supervisees are also vulnerable to misuse of power by their supervisors (Ellis et al., 2014). Researchers who study clinical supervision (e.g. Ellis et al., 2014; Murphy & Wright, 2005; Nelson et al., 2006; Szymanski, 2003, 2005) have noted that supervisors

should routinely engage their supervisees in conversations about power differentials (e.g. multiple relationships, boundary issues, responsibilities in supervision). By leveling power differentials appropriately, supervisors can create an affirming and collaborative environment (Porter & Vasquez, 1997).

Until recently, no instrument existed to objectively measure supervisees' perceptions of power in the supervisory relationship. The Power Dynamics in Supervision Scale (PDSS) is a 16-item instrument created to operationalize power dynamics in clinical supervision and to measure where supervisees perceive power lying between them and their supervisor (Cook et.al, 2018). The Cook et.al (2018) made three assumptions about power dynamics in supervision in creating the PDSS. First, power may shift between supervisors and supervisees as power in supervision is fluid. In a deeply professional and personal relationship like a supervisory relationship, issues of conflict between supervisors and supervisees may arise. These issues can be swiftly resolved or they may linger (Moskowitz & Rupert, 1983). Ongoing, unresolved conflict could trigger a supervisee to exercise their power by withholding information from the supervisor (Constantine & Sue, 2007; Murphy & Wright, 2005). Whether conflict is resolved quickly or it continues, it is how this conflict is attended to, like many other situations in supervision (e.g. providing or receiving critical feedback; McKibben et.al, 2019), that is likely to affect the power dynamics in the supervisory relationship.

Second, supervisees or supervisors may hold more or less power for some aspects of supervision, or they may share power in other areas. For example, supervisors may lead (i.e. utilize power) aspects of the supervisory process when working with novice supervisees as compared to more advanced supervisees. Novice supervisees may need more assistance from their supervisors with tasks such as case conceptualization and goal setting (Loganbill, Hardy, & Delworth, 1982), while advanced supervisees might be more self-aware of their developmental needs and better able to articulate those needs to their supervisors (Cook & Sackett, 2018). Relatedly, as conceptualized by the Discrimination model (Bernard, 1997), supervisors may be called to fulfill multiple roles (i.e. teacher, counselor, consultant), which could, in turn, change the power dynamics in the supervisory relationship depending on the role the supervisor is fulfilling in a given session. For example, supervisors who are in the role of teacher are likely to possess more control (i.e. power) of that session compared to their taking the role of a counselor or consultant.

Third, power that lies more or less with supervisors or supervisees, or power that is shared between them, is not fundamentally positive or negative. Murphy and Wright (2005) found that supervisees perceived their supervisors used power positively (e.g. sharing ideas, providing feedback, promoting safety) and negatively (e.g. imposing clinical style, abusing power, violating supervisees confidentiality). Supervisees in this study also stated that they themselves used power positively (e.g. to learn from and provide feedback to their supervisors) and negatively (e.g. to withhold information from their supervisor) as supervisees. As such, we suggest that power is not inherently helpful or problematic; rather it is how parties utilize the power in supervision that is most important. Notably, however, supervisees may be more vulnerable to abuses of power by supervisors (Ellis et al., 2014).

In the initial validation study of the PDSS, Cook and colleagues empirically tested a conceptual framework of power dynamics in clinical supervision in a sample of 267 trainees from psychology and counselor education programs. The items of the PDSS were informed by research from feminist supervision scholars (e.g. Nelson et al., 2006; Porter, 1995; Porter & Vasquez, 1997; Szymanski, 2003, 2005) as well as prior empirical research on

supervisees perceptions of power in clinical supervision (e.g. Murphy & Wright, 2005; for more information conceptual framework and item development of the PDSS, see; Cook et al. (2018). Cook et al. (2018) found that supervisees perceived themselves as possessing more power than their supervisors in maintaining healthy boundaries with their supervisors, reported a willingness to feel vulnerable in supervision, and reported feeling empowered in supervision. Supervisees perceived their supervisors as possessing more power than themselves in identifying interventions with clients, setting goals for supervision, and providing feedback about clinical skills in supervision. The creators intended the PDSS to be used as a tool to encourage a dialogue between supervisees and supervisors about the supervisees' perceptions of power dynamics in clinical supervision – thereby helping supervisors to more easily address power dynamics in supervision. Further, supervisors who choose to use the PDSS can also create a safe environment that encourages supervisees' openness and vulnerability (Szymanski, 2003). For complete information about the PDSS, including the conceptual framework, item development, and initial findings, see Cook et al. (2018).

### ***Gender, race, and power in clinical supervision***

Lacking from the initial investigation of the PDSS was an exploration into how supervisees of differing genders and races perceive power in clinical supervision. Power dynamics in supervision are inseparably linked to gender and race due to issues of privilege and oppression (Constantine & Sue, 2007; Granello, 1996; Szymanski, 2003). Feminist scholars have written extensively about gendered power imbalances in clinical supervision (Porter, 1995; Porter & Vasquez, 1997), while supervision scholars have discussed the oppression and harm endured by supervisees of Color in supervision (e.g. Constantine & Sue, 2007; Cook & Helms, 1988; Fukuyama, 1994). For example, Granello (1996) suggested that the distribution of power between supervisees and supervisors may reflect the societal gender norms to which both males and females are socialized. That is, male supervisees may be more comfortable exerting their power in the supervisory relationship, while female supervisees may be more likely to relinquish their power to their supervisors (Granello, 1996; Nelson & Holloway, 1990). Relatedly, for supervisees of Color, there is often a disparity in the distribution of power in the supervisory relationship toward the supervisor. Numerous scholars (e.g. Constantine & Sue, 2007; Cook & Helms, 1988; Fukuyama, 1994) have found that supervisees of Color felt dismissed or marginalized by White supervisors, had their clinical skills routinely questioned, and endured microaggressions in supervision. In order for supervisors to adequately attend to power, supervisors must consider how the social norms related to gender and race contribute to the power dynamics in supervision (Hernández & McDowell, 2010; Phillips, Parent, Dozier, & Jackson, 2016).

From prior research into the experiences of supervisees of differing genders and races, we can glean specific illustrations of gender- and racial-based differences in supervisees' experiences of power in clinical supervision (e.g. Constantine & Sue, 2007; Granello, Beamish, & Davis, 1997; Nelson & Holloway, 1990). For example, Granello et al. (1997) found that female supervisees as compared to male supervisees were encouraged less often by their male and female supervisors to provide input during supervision. Relatedly, Constantine and Sue (2007) found that Black supervisees perceived that their White supervisors were reluctant to give them critical feedback about their clinical skills for

fear of being perceived as a racist. Although we acknowledge that male and White supervisees may also perceive that they receive inadequate feedback, the findings from Granello et al. (1997) and Constantine and Sue (2007) highlight that female supervisees and supervisees of Color may experience power differentials in their supervisory relationship and those differentials are directly related to gender and/or race. These supervisors may have unknowingly stifled their female supervisees' and supervisees of Colors' professional development (Constantine & Sue, 2007; Granello et al., 1997; Nelson & Holloway, 1990) and violated their due process to a fair evaluation (Bernard & Goodyear, 2019; Constantine & Sue, 2007).

Relatedly, although a supervisee's ability to speak openly and to be vulnerable is a key feature of power dynamics in supervision (Cook et al., 2018; Szymanski, 2003), as compared to male supervisees, female supervisees are more guarded with what they share with their supervisors (Heru, Strong, Price, & Recupero, 2004). Similarly, supervisees of Color may hesitant to disclose information to their White supervisors, who fail to adequately attend to issues of culture in their supervisory relationship (Constantine & Sue, 2007; Cook & Helms, 1988). In sum, the findings referenced above suggest that female supervisees may perceive that they possess less power in the supervisory relationship as compared to their male peers. The same might also be true for supervisees of Color in relation to their White peers.

It should be noted that Fukuyama (1994) also found that supervisees of Color had positive experiences in cross-racial supervision, which could suggest perceived positive uses of power. These instances included receiving support from supervisors, discussing culture, and learning through multicultural activities. These results provide evidence of the importance of discussing culture and social norms as well as engaging in a collaborative partnership, which is critical to properly attend to power dynamics in clinical supervision (Szymanski, 2003).

### ***The current study***

In the absence of an instrument that operationalized supervisees' perceptions of power (i.e. the PDSS), an empirical investigation into influence of gender and race on perceived power dynamics could not be previously completed. As such, the aim of the current study was to examine if self-identified and dichotomized gender (female supervisees or male supervisees) and race membership (White supervisees and supervisees of Color) yielded differences in supervisee perceived power in clinical supervision at an overall level and specific to individual aspects of power dynamics; we discuss the implications of this dichotomization later in the manuscript. With instruments such as the PDSS, it is possible to quantitatively compare supervisees' ratings in order to evaluate the extent to which subgroups of participants and individual participants report different perceived differences in power in clinical supervision.

Table 1 provides an overview of three analytic techniques that provide insight into differences in participant responses to the PDSS. Traditionally, supervision scholars use the first approach (a. Subgroup comparisons) to compare the overall magnitude of responses (e.g. average ratings and/or subscale scores) between subgroups, which leads to broad and general implications. Although these subgroup comparisons are important, they fail to capture how subgroups or individual participants might perceive a given phenomenon differently. In addition to examining overall group differences, it is

Table 1. Summary of research questions and analytic techniques.

Analysis	Interaction Investigated	Description of Analysis	Research Question for this Study
a. Subgroup comparisons	All item responses * Subgroups	Compare the overall magnitude of responses to all items between subgroups	To what extent are there systematic differences in supervisees' perceptions of power dynamics related to gender subgroup membership (female and male), and race subgroup membership (white and person of color)?
b. Differential Item Functioning (DIF)	Subgroups * Items	Compare responses between subgroups specific to individual items	To what extent is there evidence of differential item functioning related to supervisee gender and race subgroup membership?
c. Person fit	Individual Participants * Items	Compare the progression of power imbalance for individual participants to the expected progression based on the complete sample	To what extent do individual participants' responses reflect a consistent interpretation of power dynamics with the complete sample?

also possible to conduct analyses at the level of individual items in order to better understand the development of power imbalances, and to evaluate the extent to which the nature of power imbalance is consistent for all participants.

Specifically, *differential item functioning* (DIF) analyses and *person fit analyses* based on item response theory models provide insight into interactions between responses to survey items and participant characteristics. In the context of training in psychology, Byrne et al. (2009) observed that “item response theory provides an excellent statistical framework for identifying test items that may be unfair or differentially valid across comparison groups as a consequence of gender, ethnicity, level of education, religion, culture, and other background variables” (p. 99). As shown in Table 1 part b, DIF analyses help analysts examine interactions between subgroups and items. In the context of the PDSS, items would exhibit DIF if supervisees with the same overall level of perceived power dynamics, who are members of different subgroups (e.g. female and male), would be expected to provide different ratings to the item (Smith, 1996). Whereas DIF analyses focus consistent item responses between subgroups, *person fit analyses* (Table 1 part c) focus on individual participants. In the context of the PDSS, person fit analyses allow researchers and practitioners to identify individual participants for whom the progression of various aspects of power imbalances is different from the overall sample (person *misfit*). When person misfit occurs for an individual participant, their average rating does not have the same meaning in terms of the nature of their perceived power imbalance as other participants. Accordingly, additional consideration is needed to interpret the results from the scale for these participants. Examining person fit allows for these considerations for all participants, regardless of subgroup membership.

In the current study, we applied item response theory techniques to examine three research questions. First, we considered the extent to which there were group differences in the level of reported power imbalances over all of the items: (a) To what extent are there systematic differences in supervisees’ perceptions of power dynamics related to gender subgroup membership (female and male), and race subgroup membership (White and person of Color)? This question is similar to examining main effects in analysis of variance models. Second, we examined the extent to which participants from different subgroups who had the same overall level of perceived power imbalance would be expected to respond consistently to the individual items: (b) To what extent is there evidence of differential item functioning related to supervisee gender and race subgroup membership? This question is examining interactions between subgroups and individual items. Whereas the first research question considers differences in overall levels of perceived power dynamics, this question considers how subgroups respond differently to individual items. Third, we examined the extent to which PDSS could be interpreted consistently across individual participants: (c) To what extent do individual participants’ responses reflect a consistent interpretation of power dynamics with the complete sample? This question goes beyond the first and second research questions and examines interactions between individual participants and individual items in the PDSS. Informed by prior research (e.g. Constantine & Sue, 2007; Granello et al., 1997), we hypothesized that, overall, female supervisees, as compared to their male peers, would perceive themselves as possessing less power than their supervisors. Further, we hypothesized that supervisees of Color, as compared to White supervisees, would perceive themselves as possessing less power in their supervisory relationships in relation to their supervisors.

## Method

### *Participants and recruitment procedure*

Data utilized for the current study were taken from Cook et al. (2018)'s initial survey development sample. Cook et al. (2018a) recruited master's and doctoral-level trainees who were currently receiving supervision in counseling psychology, school psychology, and counselor education programs.<sup>1</sup> The total number of respondents recruited in the study was 267; however, 38 participants did not provide demographic information related to race and/or gender, which resulted in a final sample of 229 participants. The age of participants ranged from 21 to 59 ( $M = 29.69$ ,  $SD = 7.93$ ). Most participants identified as female ( $n = 193$ , 84.28%) and White/Caucasian ( $n = 174$ , 75.98%). The sample included 173 counselor education trainees (75.5%), 53 counseling psychology trainees (23.14%), and four school psychology trainees (1.75%). The majority of participants were master's level trainees ( $n = 158$ , 69%), while 66 participants were doctoral level trainees (28.82%), and six were trainees in a doctorate of psychology (Psy.D.) program (2.62%). Although there are nuanced professional differences between counseling psychology, school psychology, and counselor education, which may have influenced their perceptions of power dynamics in clinical supervision (Cook et al., 2018), per accreditation standards, all participants were required to receive weekly supervision during an onsite field placement experience during their graduate training. The required length of the field placement experiences ranged from 640 hours (i.e. practicum and internship) for supervisees enrolled in counselor education programs (CACREP, 2015) to 1,500 to 2,000 hours for psychology students, which equates to a one to two year-long internship (APA, 2006). Full demographic information for gender and race are included in the data analysis section.

### *Instrument*

#### *PDSS (Cook et.al, 2018)*

The PDSS (see Appendix A) is a 16-item instrument designed to measure supervisees' perceptions of power in clinical supervision. Each item contains two dichotomous statements to which participants respond using a visual analog scale (VAS) to denote where they perceive power lying between them and their supervisors. Item statements on the left anchor of the VAS reflect power held by the supervisee (scored as a one), and item statements on the right anchor of the VAS reflect power held by the supervisor (scored as a four). Participants also have the option of selecting *not applicable* for each item. The PDSS is scored 1 to 4, and all not applicable responses are treated as missing data in the analysis (scored as 0). Sample stems include: (a) "I identified the goals for this supervision session" (supervisee power) to "My supervisor identified the goals for this supervision session" (supervisor power), and (b) "I trust my supervisor to keep what was discussed in this supervision session confidential" (supervisee power) to "I do not trust my supervisor to keep what was discussed in this supervision session confidential" (supervisor power). The 16 items had acceptable psychometric properties according to the Rating Scale model (Andrich, 1978) including evidence of overall model fit, person fit, and item fit (for more information, see Cook et al., 2018).



## Data analysis

We used the Rasch Rating Scale (RS) model (Andrich, 1978) to examine supervisee responses to the PDSS for evidence of subgroup differences and differential item functioning (DIF) related to supervisee subgroup membership and person fit (discussed further below). Our analyses proceeded as follows. First, we used the Winsteps software (Linacre, 2016) to estimate supervisee and item locations along a continuum that reflects perceived power dynamics between supervisees and supervisors. The continuum is constructed such that supervisees with higher locations perceived their supervisor as holding more power, and low supervisee locations indicate that a supervisee perceived themselves as holding more power. Likewise, high item locations indicate that supervisees perceived their supervisor as holding more power when responding to the item, and low item locations indicated that supervisees perceived themselves as holding more power when responding to the item. Stated mathematically, the model is as follows:  $\ln[P_{ni}(x = k)/P_{ni}(x = k - 1)] = \theta_n - \delta_i - \tau_k$ , where  $\theta_n$  is supervisee  $n$ 's estimated location on the continuum of power dynamics between supervisees and supervisors,  $\delta_i$  is the estimated location of item  $i$  on the continuum of power dynamics between supervisees and supervisors, and  $\tau_k$  is the level of perceived power dynamics associated with providing a rating in a given category.

As described in Cook et al. (2018), the RS model was an appropriate technique with which to calibrate the PDSS using our sample of  $N = 229$  participants, which exceeds Linacre's (1994) recommended sample size of around 150 participants for establishing 99% confidence intervals around Rasch model estimates within  $\frac{1}{2}$  logit.

## Differential item functioning analysis

The DIF analyses we conducted are based on the RS model. Specifically, a popular method for examining DIF within the context of the RS model is to estimate item locations separately for subgroups of interest, and calculating the absolute value of the difference in the item estimates.<sup>2</sup>

In order to conduct DIF analyses, it was necessary to group supervisees according to their self-reported demographic characteristics. Regarding gender, 193 participants identified as female (84.27%), while 32 participants identified as male (13.97%), five participants identified a gender identity other than male or female (i.e. androgynous, heterosexual, post-gender male; 2.18%). Given the low representation of non-binary participants, we excluded these participants from the analysis in order to prevent potentially inappropriate interpretation of results for this population. Accordingly, our gender-related DIF analyses involved comparisons of item locations between participants who identified as female and participants who identified as male. It should be noted that although the male subgroup was relatively small ( $n = 32$ ), the sample size for this group is within Linacre's (1994) recommended range of 27–61 participants for 99% confidence around item estimates within one logit.

Participants identified their race as: White/Caucasian ( $n = 174$ , 75.7%), Multiracial ( $n = 18$ , 7.3%), Hispanic/Latino ( $n = 13$ , 5.7%), African-American ( $n = 11$ , 4.8%), Asian ( $n = 8$ , 3.5%); American or European descent ( $n = 3$ , 1.3%), American Indian ( $n = 1$ , .4%), and Middle Eastern ( $n = 1$ , .4%). For some of these racial identity subgroups, the cell size

was too small to support meaningful comparisons using DIF analyses. The cell sizes were also too small to support using matched sampling procedures. In order to conduct DIF analyses with justifiable interpretations, we chose to dichotomize this variable to *White supervisees* and *supervisees of Color*, to balance the sample size, as consistent with prior research (e.g. Imel et al., 2011). As noted above, this grouping also allowed us to conduct DIF analyses following sample size guidelines proposed by Linacre (1994).

After we estimated supervisee and item locations, we examined differences in responses related to supervisee subgroups. First, we compared the overall magnitude of supervisee responses (average supervisee locations) between the subgroups of interest: supervisee gender subgroup membership (female or male) and race (White or person of Color). Specifically, we compared supervisees' average locations within subgroups using a two-sample *t*-test and a standardized mean difference effect size (standardized mean difference (Cohen's *d*; Cohen, 1969). We used a pooled standard deviation to calculate the standardized mean difference (Olejnik & Algina, 2000). This analysis allowed us to examine the extent to which there were systematic differences in supervisees' perceptions of power dynamics related to gender and race subgroup membership.

Finally, we conducted a DIF analysis to evaluate the degree to which there were interactions between item locations and supervisees' subgroup membership – in other words, the degree to which the magnitude of power imbalance captured by each item was systematically different depending on supervisees' subgroup membership. We used the DIF analysis procedures in the Winsteps software (Linacre, 2016) to calculate item location estimates separately for each of the subgroups that we compared in this study (male and female, white or person of color). Then, we compared these item estimates between subgroups for evidence of DIF: *Difference in item difficulty* =  $\delta_{i1} - \delta_{i2}$ , where  $\delta_{i1}$  is the item location estimate for subgroup 1, and  $\delta_{i2}$  is the item location estimate for subgroup 2. Based on this equation, a positive difference between item locations would indicate that supervisees in subgroup 1 perceived their supervisors as holding more power compared to the supervisees in subgroup 2. On the other hand, negative differences in item locations indicate that the supervisees in subgroup 1 perceived themselves as holding more power compared to the supervisees in subgroup 2. We focused on the difference in item locations between subgroups for our interpretation. Specifically, we interpreted the differences in item locations as a continuous variable, while recognizing the critical values that other researchers have proposed for classifying DIF. For example, Linacre (2016) provided guidelines for interpreting differences in Rasch item difficulty estimates that reflect other DIF approaches. Linacre recommended that researchers interpret item difficulty differences as follows: *difference* < 0.43 logits = negligible DIF, 0.43 logits ≤ *difference* ≤ 0.63 logits = slight-to-moderate DIF, and *difference* ≥ 0.64 logits = substantial DIF.

We also evaluated these differences in item difficulty using a paired *t*-test:

$$t_i = \delta_{i1} - \delta_{i1} / s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (1)$$

In Equation 1,  $s_p$  is the pooled variance:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2) - 2}} \quad (2)$$

$\delta_{i1}$  is the location of item  $i$  estimated for supervisee subgroup 1,  $\delta_{i2}$  is the location of item  $i$  estimated for supervisee subgroup 2,  $SE_{i1}$  is the standard error of estimate for  $\delta_{i1}$ ,  $SE_{i2}$  is the standard error of estimate for  $\delta_{i2}$ ,  $n_1$  is the number of supervisees in subgroup 1, and  $n_2$  is the number of supervisees in subpopulation 2 (Smith, 1996). This  $t$ -test allowed us to evaluate the statistical significance of the difference in item locations between pairs of subgroups.

### Individual supervisee fit

Although supervisees' gender and racial identities impact their experiences in clinical supervision in important ways, we recognize that grouping procedures in general and our dichotomized grouping scheme in particular, provides a limited view of individual supervisees' experiences with power dynamics in these settings. Accordingly, we examined supervisees' responses to the PDSS using person fit analyses<sup>3</sup> (Glas & Khalid, 2016; Meijer & Sijtsma, 2001; Smith, 1986) as a supplement to our DIF analyses. In the context of this study, the supervisees are the persons who responded to the PDSS, so we will refer to these analyses as *supervisee fit analyses* from now on. Supervisee fit analyses are methods for evaluating the hypothesis that the PDSS measures individual supervisee locations on the perceptions of power dynamics construct in a meaningful way. It is reasonable to interpret a supervisee's location estimate when there is evidence suggesting that the supervisee interpreted the progression of power imbalances between supervisees and supervisors (i.e. relative item ordering) in a similar way as the overall sample. When this assumption is not reasonably met for individual supervisees, it is not appropriate to interpret their location estimate. Individual supervisee fit analyses allow us to consider the psychometric properties of the PDSS for individuals without subgroup comparisons and the corresponding limitations due to combining subgroups.

Numerous methods are available for evaluating person fit on scales such as the PDSS, including a variety of numeric and graphical indicators (Walker, Jennings, & Engelhard, 2018). Recognizing the limitations associated with using a single person fit indicator, we used a combination of numeric fit statistics based on the measurement theory that we used to guide the development of the PDSS, and we supplemented our interpretation of these statistics with a graphical indicator of person fit.

### Numeric fit statistics

We selected two person fit statistics for our analysis: outfit mean square error ( $MSE$ ) and infit  $MSE$ . We selected these statistics because they are designed to detect fit to the measurement model that we used to scale the PDSS (the RS model). In the context of the PDSS, these statistics are averages of *supervisee residuals*, or discrepancies between the response that a supervisee provided on an item and the response that would be expected given their location ( $\theta$ ). Outfit  $MSE$  is calculated as follows:

$$\text{Outfit } MSE = \sum_i^L z_{ni}^2 / L \quad (3)$$

where  $Z_{ni}$  is the standardized residual between the observed and expected response for Supervisee  $n$  on item  $i$ , and  $L$  is the number of items. Outfit  $MSE$  is sensitive to extreme unexpected supervisee responses, so it can alert analysts to supervisees whose response patterns include extreme discrepancies from model expectations. On the other hand, infit  $MSE$  is a weighted average statistic, so it is less sensitive to extreme residuals:

$$\text{Infit } MSE = \sum_i^L y_{ni}^2 / \sum_i^L W_{ni}^2 \quad (4)$$

In the Infit  $MSE$  statistic, the numerator includes standardized residuals ( $y_{ni}$ ) that are weighted by the variance of a person's response pattern ( $W_{ni}$ ). Because the infit statistic weights the standardized residuals by the variance, the residuals closest to a supervisee's location estimate have the most influence. For both outfit  $MSE$  and infit  $MSE$ , higher values indicate more frequent and extreme residuals than expected by the probabilistic model, and lower values indicate less frequent and extreme residuals than expected (i.e. overly consistent responses).

Numeric fit statistics are continuous variables, and they are sensitive to a variety of factors, including the particular sample from which they are estimated (Smith, 2004; Wolfe, 2013). Following Wolfe (2013) and Walker et al. (2018), we used a nonparametric bootstrap procedure to identify critical values that reflect the empirical distribution of person fit in our data. Specifically, we used the supervisee and item parameters from our original RS model estimates to generate 1,000 datasets and analyzed them using the RS model. From these datasets, we calculated person fit statistics for each supervisee, and identified the value of the person fit statistic at the 97.5<sup>th</sup> percentile as our critical value to identify supervisee misfit. We selected this value to reflect the upper tail of a normal probability distribution in which values that exceed the top 2.5<sup>th</sup> percentile or the bottom 2.5<sup>th</sup> percentile are considered statistically significant (e.g.  $p < 0.05$ ). Higher-than-expected person fit statistics indicate large deviations between the responses that would be expected given model estimates. In contrast, lower-than-expected person fit statistics indicate consistency in responses. As a result, researchers typically consider higher-than-expected person fit statistics as more cause for concern than lower-than-expected person fit statistics because they indicate that the person of interest has distinguished between the items in a way that is inconsistent with the item ordering observed for the rest of the sample, such that the overall item ordering is not meaningful (Bond & Fox, 2015; Meijer, Niessen, & Tendeiro, 2016). In contrast, low person fit statistics indicate that a participant has not provided different responses to the items (e.g. they provided ratings in the same rating scale categories to most of the items). Reflecting this perspective, we used the upper tail of the distribution as the critical value for identifying person misfit.

### **Graphical person fit analysis**

We used graphical analyses to supplement the numeric supervisee fit analyses. Graphical analyses provide analysts with insight into the nature of expected and unexpected supervisee responses over a set of items that is often more diagnostically useful than a classification of "fitting" or "misfitting" based on a numeric fit statistic. The graphical analysis technique that we employed was *person response functions* (PRF) for individual supervisees. A PRF displays the model-expected (i.e. theoretical) item responses, based on

the supervisee's estimated location ( $\theta$ ) over the set of items. The observed item responses are plotted in the same coordinate space as the expected responses in order to identify discrepancies. A 95% confidence interval can be included in the plot to guide the interpretation of these discrepancies.

## Results

Before we examined the PDSS results for evidence of differences in supervisee locations and DIF, we evaluated the overall fit of the responses to the Rating Scale model. This analysis helped us to ensure that the instrument exhibited generally acceptable psychometric properties such that we could meaningfully interpret group differences in responses to the PDSS. As we observed in Cook et al. (2018), our sample of supervisees' responses to the 16-item scale demonstrated acceptable overall fit to the model, with 60.6% of the variance explained by the Rasch measures. This value is well above the 20% critical value suggested by Reckase (1979) for Rasch analyses of potentially multidimensional scales. This result suggests that the supervisees' perceptions of power were the main contributing factor in their responses to the PDSS items. For additional details about overall model fit and individual item estimates and item fit, see Cook et al. (2018).

In the following paragraphs, we present results related to the three research questions, as addressed by the three analytic techniques presented in Table 1. A discussion of the results follows.

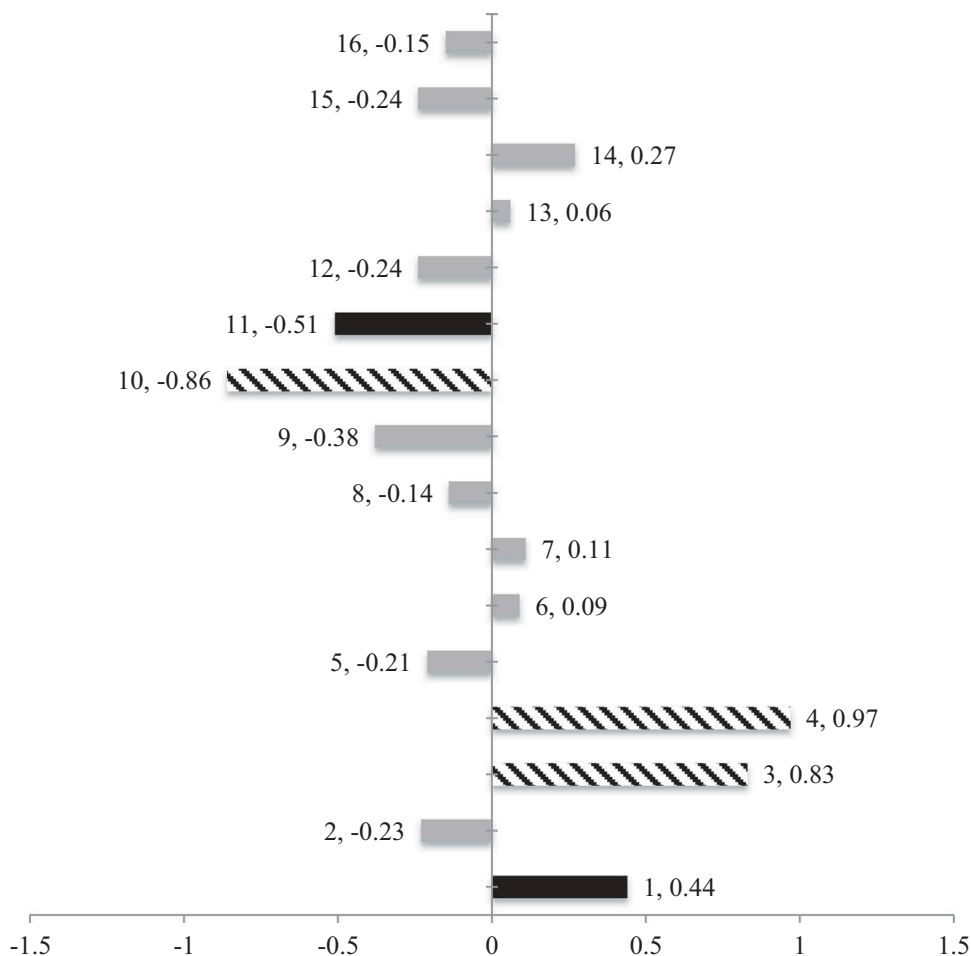
### *a. Differences in supervisee locations*

Using supervisee estimates from the Rating Scale model, we compared the overall magnitude of reported power imbalances between supervisee subgroups related to gender (female supervisees and male supervisees) and race (White supervisees or supervisees of Color) subgroups over all of the items. To do this, we used the average estimated locations on the power dynamics construct between subgroups. Contrary to our hypothesis, the results from these analyses did not indicate systematic differences in power dynamics between gender subgroups ( $t(40.37) = 0.52, p = 0.61$ , standardized mean difference = 0.10) and race subgroups ( $t(85.19) = -0.20, p = 0.84, d = 0.03$ ). Although these non-significant differences in the overall average locations of supervisees are interesting, their interpretation is unclear without evidence related to DIF and person fit. Specifically, DIF and person fit analyses are needed to understand the extent to which the nature of power dynamics is consistent between subgroups and for individual participants.

### *b. Differential item functioning*

Next, we examined the PDSS for evidence of DIF related to supervisee gender (female or male). These analyses allowed us to investigate interactions between subgroups and individual items. Figure 1 provides a visual summary of the differences in the relative magnitude of responses to each item between subgroups. Specifically, in the figure, the difference in item responses between female supervisees and male supervisees is shown along the x-axis. Items are arranged on the y-axis in item order from the PDSS instrument.

### Item Difficulty Comparison between Female and Male Supervisees



**Figure 1.** Differences between item difficulty calibration for female and male supervisees.

Notes. (1). The comparisons were calculated as  $\delta_{ifemale} - \delta_{imale}$ , such that bars that point to the right of the graph indicate that female supervisees reported that their supervisors held more power more often than male supervisees. Bars that point to the left of the graph indicate that male supervisees reported that their supervisors held more power more often than female supervisees. (2) Labels at the end of each bar indicate the item identification number, followed by the value of the difference between item difficulty locations ( $\delta_{ifemale} - \delta_{imale}$ ). (3) Shading is used to indicate the magnitude of DIF as follows: light gray bar = negligible DIF ( $\delta_{ifemale} - \delta_{imale} < 0.43$  logits), dark black bar = slight-to-moderate DIF ( $0.43 \text{ logits} \leq \delta_{ifemale} - \delta_{imale} \leq 0.63 \text{ logits}$ ), and diagonal striped bar = substantial DIF ( $\delta_{ifemale} - \delta_{imale} \geq 0.64 \text{ logits}$ ).

Shaded bars show the difference in item estimates between female supervisees and male supervisees specific to the individual item. Bars that point to the right of the graph indicate that female supervisees reported that their supervisors held more power more often than male supervisees related to the specific item. Conversely, bars that point to the left of the graph indicate that male supervisees reported that their supervisors held more power more often than female supervisees, specific to the item of interest. At the end of each bar, a label is provided that shows the item identification number, followed by the

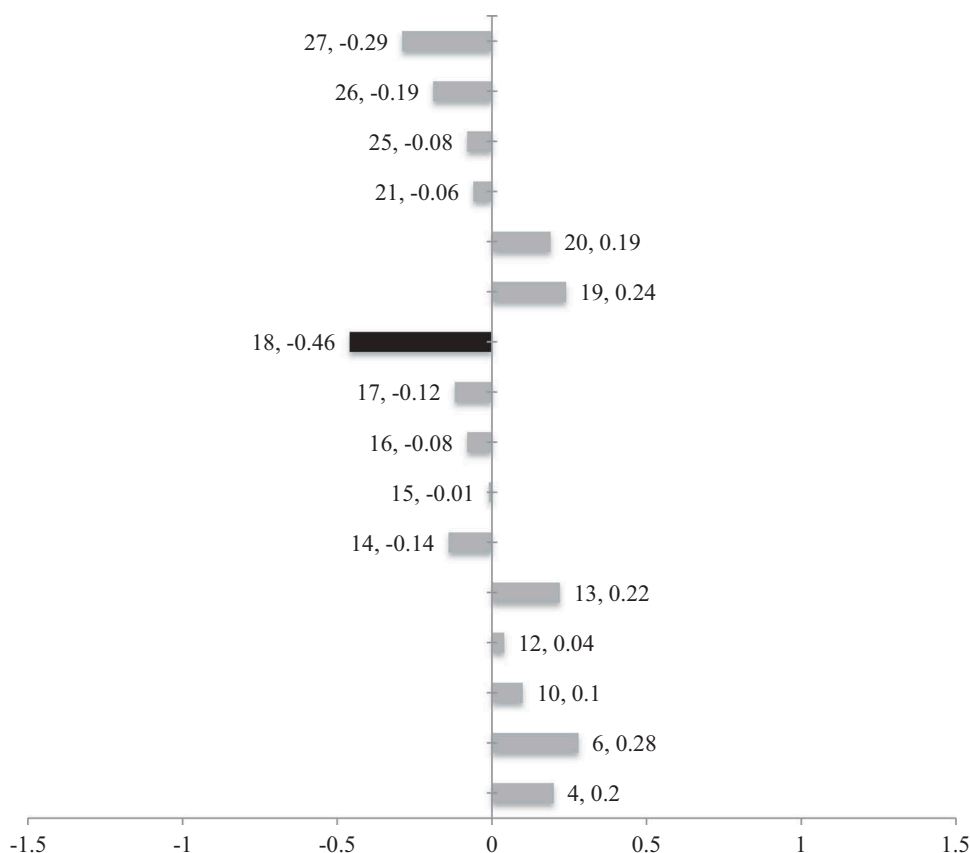
value of the difference between item locations ( $\delta_{ifemale} - \delta_{imale}$ ). We used Linacre's (2016) classifications to shade the bars according to the magnitude of DIF, where lightly shaded solid bars indicate negligible DIF, dark shaded solid bars indicate slight-to-moderate DIF, and diagonal striped bars indicate substantial DIF. Across the 16 items, two items exhibited slight-to-moderate DIF: Female supervisees reported more power imbalance in favor of their supervisors on Item 1 (i.e. "identifying goals for supervision") ( $\delta_{ifemale} - \delta_{imale} = 0.44$ ), and male supervisees reported more power imbalance in favor of their supervisors on Item 11 (i.e. "feedback on clinical skills") ( $\delta_{ifemale} - \delta_{imale} = -0.51$ ). The absolute value of the difference between Item 1 and Item 11 was 0.07 logits, which indicates that although the two items exhibited DIF in opposite directions, the magnitude of the differences were similar for the two items. Three items exhibited statistically significant and substantial DIF: Female supervisees reported more power imbalance in favor of their supervisors on Item 3 (i.e. "conceptualizing client cases") ( $\delta_{ifemale} - \delta_{imale} = 0.83, p < 0.05$ ) and Item 4 (i.e. "facilitating discussions of power") ( $\delta_{ifemale} - \delta_{imale} = 0.97, p < 0.05$ ), and male supervisees reported more power imbalance in favor of their supervisors on Item 18 (i.e. "feedback on work with clients") ( $\delta_{ifemale} - \delta_{imale} = -0.86, p < 0.05$ ).

Figure 2 shows the DIF results related to supervisee race (white or person of color), using the same format as Figure 1. We calculated the item difficulty estimates by subtracting the item estimate for White supervisees from the item estimate for supervisees of Color ( $\delta_{iperson\ of\ color} - \delta_{iwhite}$ ), such that positive differences (bars pointing to the right) indicate that supervisees of Color reported that their supervisors held more power more often than white supervisees, and negative differences (bars pointing to the left) indicate that white supervisees reported that their supervisors held more power more often than supervisees of Color. We observed one item with slight-to-moderate DIF related to supervisee race: White supervisees reported more power imbalance in favor of their supervisors on Item 10 (i.e. "feedback on work with clients") ( $\delta_{iperson\ of\ color} - \delta_{iwhite} = -0.46, p = 0.36$ ). However, none of the differences in item difficulty calibrations related to supervisee race were statistically significant.

### Individual supervisee fit

The critical values for identifying person misfit based on our empirical nonparametric bootstrap procedure were equal to 1.895 and 1.708 for outfit *MSE* and infit *MSE*, respectively. Using both of these criteria, 91% of the supervisees' response patterns adequately fit their location estimates based on the RS model. To support the interpretation of these results, we examined PRFs for supervisees who fit and did not fit the model according to these critical values. Overall, the PRFs supported the conclusions that would be made about supervisee fit based on the numeric fit statistics. To illustrate these findings, Figure 3 includes PRFs for two supervisees with adequate person fit (outfit *MSE*  $\leq 1.895$  and infit *MSE*  $\leq 1.708$ ) and two supervisees with inadequate person fit (outfit *MSE*  $> 1.895$  and infit *MSE*  $> 1.708$ ). In each plot, the x-axis shows the PDSS items ordered such that participant responses would be expected to be non-increasing, given the overall item ordering from the complete sample. The y-axis shows the PDSS rating scale. Solid triangle plotting symbols with a solid line show the model-expected responses, and open circle plotting symbols with a dashed line show the observed responses. Thin lines show a 95% confidence interval around the expected

### Item Difficulty Comparison between Person of Color and White Supervisees



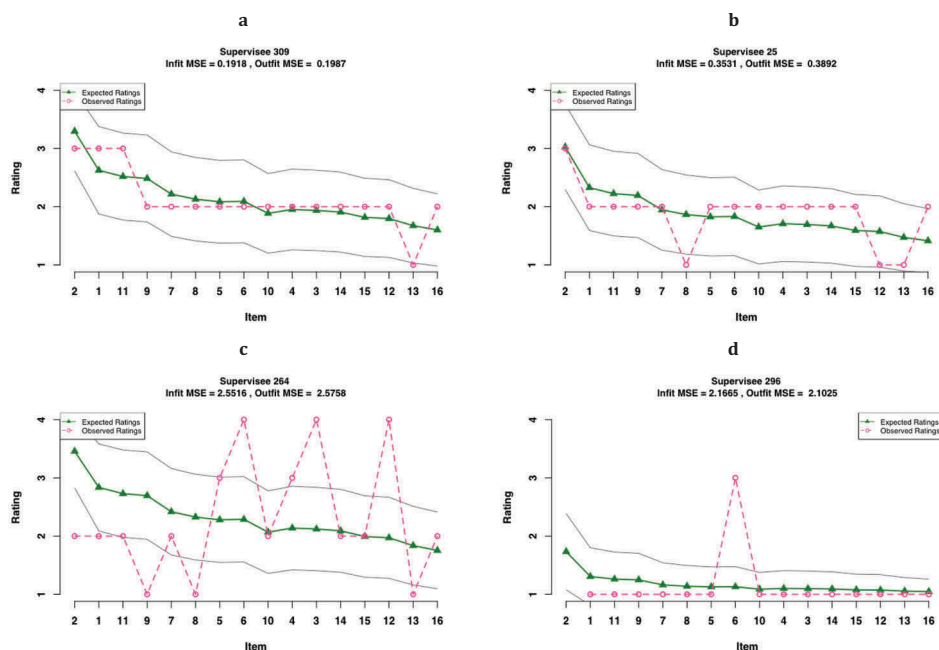
**Figure 2.** Differences between item difficulty calibration for white and person of color supervisees.

Notes. (1). The comparisons were calculated as  $\delta_{iwhite} - \delta_{iperson\ of\ color}$ , such that bars that point to the left of the graph indicate that white supervisees reported that their supervisors held more power more often than supervisees of Color. Bars that point to the right of the graph indicate that supervisees of Color reported that their supervisors held more power more often than white supervisees. (2) Labels at the end of each bar indicate the item identification number, followed by the value of the difference between item difficulty locations ( $\delta_{iwhite} - \delta_{iperson\ of\ color}$ ). (3) Shading is used to indicate the magnitude of DIF as follows: light gray bar = negligible DIF ( $\delta_{iwhite} - \delta_{iperson\ of\ color} < 0.43$  logits), dark black bar = slight-to-moderate DIF ( $0.43 \text{ logits} \leq \delta_{iwhite} - \delta_{iperson\ of\ color} \leq 0.63$  logits), and diagonal striped bar = substantial DIF ( $\delta_{iwhite} - \delta_{iperson\ of\ color} \geq 0.64$  logits).

response function. If supervisees' responses fit the model, we would expect to see non-increasing ratings as items progress from left to right on the x-axis.

Supervisee 309 (Figure 3 (a)), who identified as White and Female, is an example of a supervisee whose responses matched well with the pattern that would be expected given their location estimate (infit  $MSE = 0.19$ ; outfit  $MSE = 0.20$ ). Although this supervisee gave a lower-than-expected response to item 13, the overall pattern of their responses suggests that they interpreted the relative ordering of the PDSS items in generally the same way as the overall sample. These results support the interpretation of this supervisee's location estimate as an indicator of their perception of power dynamics. The results





**Figure 3.** Observed and Expected Person Response Functions for Selected Supervisees.

for Supervisee 25, who identified as White and Male, (Figure 3 (b)); infit  $MSE = 0.35$ , outfit  $MSE = 0.40$ ) provided similar evidence.

The remaining plots (Figure 3(c and d)) show PRFs for two examples of supervisees whose numeric fit statistics suggested that their responses did not adequately match model expectations. For Supervisee 264 (Figure 3 (c)); infit  $MSE = 2.55$ , outfit  $MSE = 2.58$ ), who identified as White and Female, there were many substantial discrepancies between the observed and expected PRFs. For example, this supervisee gave lower-than-expected ratings to items for which the overall sample gave relatively high ratings, such as Item 9 (*I trust my supervisor to keep what was discussed in this supervision session confidential*) and Item 8 (*I was able to speak freely in this supervision session*). Likewise, on items for which the overall sample provided generally low ratings, this supervisee gave relatively high ratings: Item 6 (*The evaluation of my work that I received in this supervision session benefited me as a counselor*), Item 3 (*I conceptualized my client cases in this supervision session*), and Item 12 (*I felt like I could be vulnerable in this supervision session with my supervisor*). As a result of these frequent unexpected ratings, the overall shape of the observed PRF was notably different from that of the expected PRF. Together, the numeric and graphical person fit analyses do not support the interpretation of the location estimate for Supervisee 264.

Another example of a supervisee whose numeric fit statistics indicated misfit to the RS model is Supervisee 296, who identified as Female and African American (Figure 3 (d)); infit  $MSE = 2.17$ , outfit  $MSE = 2.10$ ). This supervisee's expected and observed PRFs show an interesting pattern. Specifically, all of Supervisee 296's responses were within the range that would be expected given their estimated location on the power dynamics construct

except for item 6 (*The evaluation of my work that I received in this supervision session benefited me as a counselor*). For this item, Supervisee 296 gave a higher rating than expected in light of their other responses and in light of the ordering for the PDSS items based on the rest of the participants' responses – given these other values, the model-expected rating for Supervisee 296 on Item 6 was equal to 1, but the observed rating was equal to 3. This substantial deviation from the expected rating resulted in person misfit for Supervisee 296. Although there was only one notable deviation from model expectations, the unexpected response to item 6 warrants additional consideration before this supervisee's location estimate can be meaningfully interpreted.

## Discussion

The aim of the current study was to examine the extent to which gender (self-identified and dichotomized) and race yielded differences in supervisee perceived power in clinical supervision. With the current sample, we found that the perceptions of power dynamics in clinical supervision were not meaningfully different between subgroups defined using gender (male or female) or race (White or Supervisees of color). These findings are important because they seem to contrast conceptual understandings of power dynamics in supervision. Notably, effect sizes for gender and race were very small, indicating that other factors may better explain overall perceptual differences in power. For example, in this study, we did not control for variables such as the supervisory working alliance or the supervisor's approach. McKibben, Cook, and Fickling (2019) found that when supervisees perceive their supervisor to be utilizing a feminist approach to supervision, the working alliance tends to be stronger and supervisees are less likely to intentionally withhold information. It is possible that supervisees throughout our sample, regardless of race or gender, experienced supervisors who were culturally appropriate and responsive, which served to mitigate power differentials. Another possibility, which we discuss further below as a limitation to the study, is that the majority White, female sample limited our ability to observe nuanced differences among race and gender subgroups.

Power dynamics are nuanced (Cook et al., 2018), as are cultural dynamics, and we also found evidence for gender- and race-based power differentials for specific features of power in clinical supervision, which are also important to also interpret. Female supervisees perceived their supervisors as possessing more power in areas of clinical supervision that included establishing goals in supervision and conceptualizing client cases, while male supervisees perceived their supervisors as possessing more power providing feedback about their clients and clinical skills. Since the majority of participants were master's students, and because neophyte trainees tend to need more guidance to mitigate anxiety, learn skills, and navigate the therapeutic process with clients (Stoltenberg & McNeill, 2010), one might expect for a supervisor to more often fulfill the function of a teacher (Bernard, 1997). This may explain why male and female supervisees perceived their supervisors as possessing power in key of clinical supervision (i.e. goal setting, case conceptualization, giving feedback).

It is also important to consider that prior research suggests that supervisors less frequently solicit female supervisee input in clinical supervision (e.g. Granello et al., 1997; Nelson & Holloway, 1990). Further, female supervisees may be less willing to disclose their concerns about the power dynamics to their clinical supervisors (Heru et al., 2004). In fact,

the female supervisees in the current study, as compared to their male peers, perceived their supervisors as possessing more power in facilitating discussions of power dynamics in their supervisory relationships. Such disparities may negatively impact female supervisees' long-term clinical skill development, and further perpetuate the idea that they are more reliant on their supervisors than their male peers (Granello, 2003). In sum, these findings suggest that supervisors may need to engage their supervisees of all genders in a conversation about their perceptions of the nuances of power in clinical supervision in order to provide opportunities for supervisees to facilitate their own autonomy (e.g. co-constructing supervision goals, contributing to client case conceptualization), while also providing developmentally appropriate supervision (e.g. giving adequate feedback about client and clinical skills; Bernard & Goodyear, 2019; Ellis et al., 2014).

Regarding race, we found a slight to moderate difference in that White supervisees perceived supervisors as more in power when providing feedback about their clients (i.e. supervisors did not provide feedback about their clients), but this difference was not statistically significant. There is evidence from prior qualitative research that supervisees of Color perceived that they received inadequate feedback from their supervisors, who they believed may have been concerned about being perceived as racist (e.g. Constantine & Sue, 2007). Our findings, which reflected the perceptions of both White supervisees and supervisees of Color, may suggest that novice supervisees (i.e. trainees) of all races may desire more feedback than their supervisors are currently providing. However, because we did not capture the participants' reasons for their perspectives, we are unable to offer inferences into this finding. Similarly, we are unable to assert whether these experiences were perceived by the supervisees in this study as negative or positive, because we neither assigned value judgment to power differentials (i.e. that a supervisor or supervisee holding more/less power in a given area is good or bad) with the PDSS, nor did we distinguish between positive and negative power dynamics in this study.

Finally, individual interpretations are often ignored in group comparisons, and we want to highlight the importance of considering individuals, especially when subgroup analyses are challenging to complete due to small sample sizes. More specifically, the person fit analysis suggested that the power dynamics are not experienced consistently for all supervisees. Since we do not assign power as being positive or negative, we cannot interpret the degree to which it is most beneficial for supervisees to perceive themselves as possessing power, nor can we interpret whether supervisees, like Supervisees 264 and 296, felt the distribution of power in their supervisors relationship was advantageous or not. However, these examples do provide evidence that supervisors may need to discuss all aspects of clinical supervision to fully understand their supervisees' perceptions of power in their relationship.

## **Limitations**

The current findings should be considered in light of several limitations. First, because of our sampling procedures we were unable to calculate a response rate. Also, we only examined the supervisees' perception of power dynamics in clinical supervision. This may yield a one-sided picture of power in clinical supervision. Relatedly, an important topic in multicultural supervision is the influence of supervisors' gender and race on the supervisory process as well as supervision outcomes in cultural matched and cross-cultural

supervision dyads (Granello, 1996). However, the participants in the current study provided demographic information about their supervisors. Given that the supervisors did not provide this demographic information themselves, we decided that any analyses of supervisees' perceptions of power based on supervisor gender or race as well as power dynamics in cross-cultural and matched dyads could be problematic. This may be particularly true for our non-significant findings regarding race. For example, our sample included an overrepresentation of White participants in the current sample (75.7%), which may have limited our ability to observe more nuanced race-based differences in perception of power. Further, the majority of previous research on race in clinical supervision examined the experiences of supervisees of Color who were receiving supervision from White supervisors. In the current study, the majority of supervisees identified as White and the majority of supervisors were identified as White (77.4%). Thus, it is possible that the results from the current study are limited to supervisees' perceptions of power in culturally matched dyads (i.e. White supervisee -White supervisor). Within-group heterogeneity may also be a factor in the current findings, suggesting that additional factors within racial groups may contribute to perceptions of power. Finally, limitation in the current study was our decision to dichotomize variables (e.g. White supervisee, supervisees of Color) and to exclude supervisees who reported a nonbinary gender. We recognize that male/female is gender-binary and may be more related to biological sex than gender. Further, this coding may not necessarily be inclusive of all gender identities. However, we allowed participants to self-identify their gender with an open-ended question in our original survey, so the male/female response are self-identifiers and not necessarily our descriptors. We also recognize the irony of combining the groups into "supervisees of Color," which means that we are forced to broadly interpret the results. Although our approach is consistent with prior studies and best practices from a statistical perspective, our findings may be limited to our coding of variables. It is also possible that further recruiting a sufficient sample of supervisees of Color might mitigate our need to collapse supervisees of Color and to detect significant effects related to supervisee race.

### ***Implications***

Feminist scholars suggest that respect for diversity should be a core feature of supervision, and directly discussing how gender and racial societal norms manifest in the supervisory relationship is critical for supervisors (Porter, 1995; Porter & Vasquez, 1997). Regarding gender, we found evidence that supervisors may not ask female supervisees to contribute to key aspects of the supervision process (i.e. case conceptualization, goal setting), while male supervisees may not receive needed feedback. Both scenarios could be damaging to female and male supervisees' personal and professional development. When supervising a female trainee, supervisors may need to be intentional to safeguard that their female supervisees are given an opportunity to set goals for supervision and to conceptualize client cases. Relatedly, male supervisees need feedback at a frequency that is appropriate given their developmental level. In sum, supervisors are urged to reflect on their potentially biased assumptions about gender norms (Hernández & McDowell, 2010) to ensure that they are providing equitable and developmentally appropriate supervision to their male and female supervisees.

Further, although we found no statistical differences in how White supervisees and supervisees of Color perceive power in clinical supervision overall, we caution supervisors from assuming that race is not a salient factor in their supervisory relationship. As evidenced by the person fit analyses for individual supervisees, supervisors cannot reasonably anticipate a consistent pattern in how all supervisees experience the power in clinical supervision. That is, supervisors cannot assume supervisees who perceives themselves as possessing a high degree of power for one aspect of clinical supervision will also perceive themselves as holding a high level of power for another aspect. Thus, supervisors may need to discuss all facets of the power the supervisory relationship with their supervisees. Such discussions may be a time-consuming task, thus the PDSS may be a useful tool to help supervisors quickly operationalize sources of power in supervision, and to begin discussions about gender and racial norms in the supervision process (Hernández & McDowell, 2010; Phillips et al., 2016).

### **Conclusion and future research**

In sum, our findings provide insights into the some of the nuanced differences in supervisees' perceptions of power dynamics in clinical supervision, while also raising more questions about the intersections of race, gender, and power in supervision. This study is an important next step into the examination of power in clinical supervision, and future researchers can build on this work to examine the influence of supervisor demographic variables as well as the power dynamics in cross-cultural and culturally matched supervision dyads. Further, future researchers are encouraged to investigate more nuanced differences within cultural groups *and* among cultural identities given that race and gender are not homogeneous. Relatedly, scholars may want to consider the influence of other cultural identities such as religion, socioeconomic status, sexual orientation, and disability status (Bernard & Goodyear, 2019). Collectively, supervision scholars can inform supervision practices and to help supervisors and supervisees better attend to the power dynamics in their supervisory relationships.

### **Notes**

1. In the initial analysis of the PDSS, Cook et al. (2018) observed some small differences in the overall level of perceptions of power dynamics between the Master's level and Doctoral level participants that indicated potential developmental differences between the two groups. However, dimensionality analyses revealed that the instrument functioned comparably between these groups. Please see Authors et al. (2018a) for more details.
2. This method is equivalent to the method that DIF method that Raju (1988) described based on the difference between item response functions (IRFs); specifically, the absolute value of the difference in item calibration is equivalent to the space between IRFs for one-parameter logistic models and Rasch models, such as the RS model (Gamerman, Goncalves, & Soares, 2018).
3. In other studies, researchers have used a variety of names for these analyses, including person reliability, appropriateness measurement, person fit, among others (Walker et al., 2018).

### **Disclosure statement**

No potential conflict of interest was reported by the authors.

## Notes on contributors

**Stefanie A. Wind**, PhD is an Assistant Professor of Educational Measurement at the University of Alabama. Her primary research interests include the exploration of methodological issues in the field of educational measurement, with emphases on methods related to rater-mediated assessments, rating scales, Rasch models and item response theory models, and nonparametric item response theory, as well as applications of these methods to substantive areas related to education.

**Ryan M. Cook**, PhD, LPC, ACS, is an Assistant Professor in the Department of Educational Studies in Psychology, Research Methodology, and Counseling at the University of Alabama. His research interests include supervisee nondisclosure, clinical supervision, and professional issues in the field of counseling.

**W. Bradley McKibben**, PhD, IMH (FL), NCC is an Assistant Professor in the Department of Counseling at Nova Southeastern University. His research interests include professional counselor development and relational/cultural issues in clinical supervision.

## ORCID

Stefanie A. Wind  <http://orcid.org/0000-0002-1599-375X>

Ryan M. Cook  <http://orcid.org/0000-0003-2661-8274>

## References

- American Psychological Association. (2006). *Guidelines and principles for accreditation of programs in professional psychology (G&P)*. Retrieved from <http://www.apa.org/ed/accreditation/about/policies/guiding-principles.pdf>
- Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Bernard, J. M. (1997). The discrimination model. In C. E. Walktins Jr. (Ed.), *Handbook of psychotherapy supervision* (pp. 31–327). New York, NY: Wiley.
- Bernard, J. M., & Goodyear, R. K. (2019). *Fundamentals of clinical supervision*. Boston, MA: Pearson.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.
- Byrne, B. M, Oakland, T, Leong, F T. L, van de Vijver, F J. R, Hambleton, R. K, Cheung, F. M, & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: implications for improved education and training in psychology. *Training and Education in Professional Psychology*, 3(2), 94–105. doi:10.1037/a0014516
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic.
- Constantine, M. G., & Sue, D. W. (2007). Perceptions of racial microaggressions among Black supervisees in cross-racial dyads. *Journal of Counseling Psychology*, 54, 142–153.
- Cook, D. A., & Helms, J. E. (1988). Visible racial/ethnic group supervisees' satisfaction with cross-cultural supervision as predicted by relationship characteristics. *Journal of Counseling Psychology*, 35(3), 268.
- Cook, R, McKibben, B, & Wind, S. A. (2018). Supervisee perception of power in clinical supervision: the power dynamics in supervision scale. *Training and Education in Professional Psychology*, 12(3), 188–195. doi:10.1037/tep0000201
- Cook, R. M, & Sackett, C. R. (2018). Exploration of prelicensed counselors experiences prioritizing information for clinical supervision. *Journal Of Counseling & Development*, 96, 449–460. doi: 10.1002/jcad.2018.96.issue-4
- Council for Accreditation of Counseling and Related Educational Programs. 2015. *2016 standards*. Alexandria, VA: Author. Retrieved from <http://www.cacrep.org/wp-content/uploads/2012/10/2016-CACREP-Standards.pdf>



- Ellis, M. V., Berger, L., Hanus, A., Ayala, E. E., Swords, B. A., & Siembor, M. J. (2014). Inadequate and harmful clinical supervision: Testing a revised framework and assessing occurrence. *The Counseling Psychologist*, 42, 434–472.
- Fukuyama, M. A. (1994). Critical Incidents in Multicultural Counseling Supervision: A Phenomenological Approach to Supervision Research. *Counselor Education and Supervision*, 34, 142–151.
- Gamerman, D., Goncalves, F. B., & Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 3, pp. 67–86). Boca Raton, FL: CRC Press.
- Glas, C. A. W., & Khalid, N. (2016). Person fit. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 3, pp. 107–126). Boca Raton, FL: CRC Press.
- Granello, D. H. (1996). Gender and power in the supervisory dyad. *The Clinical Supervisor*, 14, 53–67.
- Granello, D. H. (2003). Influence strategies in the supervisory dyad: An investigation into the effects of gender and age. *Counselor Education and Supervision*, 42(3), 189–202.
- Granello, D. H., Beamish, P. M., & Davis, T. E. (1997). Supervisee empowerment: Does gender make a difference? *Counselor Education and Supervision*, 36, 305–317.
- Hernández, P., & McDowell, T. (2010). Intersectionality, power, and relational safety in context: Key concepts in clinical supervision. *Training and Education in Professional Psychology*, 4(1), 29–35.
- Heru, A. M., Strong, D. R., Price, M., & Recupero, P. R. (2004). Boundaries in psychotherapy supervision. *American Journal of Psychotherapy*, 58(1), 76–89.
- Imel, Z. E., Baldwin, S., Atkins, D. C., Owen, J., Baardseth, T., & Wampold, B. E. (2011). Racial/ethnic disparities in therapist effectiveness: A conceptualization and initial study of cultural competence. *Journal of Counseling Psychology*, 58, 290–298.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.
- Linacre, J. M. (2016). Winsteps Rasch Measurement (Version 3.92.1) [Computer software]. Winsteps.com
- Loganbill, C., Hardy, E., & Delworth, U. (1982). Supervision: a conceptual model. *The Counseling Psychologist*, 10(1), 3–42. doi: [10.1177/0011000082101002](https://doi.org/10.1177/0011000082101002)
- McKibben, W. B., Borders, L. D., & Wahesh, E. (2019). Factors influencing supervisee perceptions of critical feedback validity. *Counselor Education and Supervision*, 58, 242–256. doi: [10.1002/ceas.v58.4](https://doi.org/10.1002/ceas.v58.4)
- McKibben, W. B., Cook, R. M., & Fickling, M. (2019). Feminist supervision and supervisee nondisclosure: the mediating role of the supervisory relationship. *The Clinical Supervisor*, 38, 38–57. doi: [10.1080/07325223.2018.1509756](https://doi.org/10.1080/07325223.2018.1509756)
- Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, 23(1), 52–62.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.
- Moskowitz, S. A., & Rupert, P. A. (1983). Conflict resolution within the supervisory relationship. *Professional Psychology: Research and Practice*, 14(5), 632–641.
- Murphy, M. J., & Wright, D. W. (2005). Supervisees' perspectives of power use in supervision. *Journal of Marital and Family Therapy*, 31, 283–295.
- Nelson, M. L., Gizara, S., Hope, A. C., Phelps, R., Steward, R., & Weitzman, L. (2006). A feminist multicultural perspective on supervision. *Journal of Multicultural Counseling and Development*, 34, 105–115.
- Nelson, M. L., & Holloway, E. L. (1990). Relation of gender to power and involvement in supervision. *Journal of Counseling Psychology*, 37, 473–481.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241–286.
- Phillips, J. C., Parent, M. C., Dozier, C., & Jackson, P. L. (2016). Depth of discussion of multicultural identities in supervision and supervisory outcomes. *Counselling Psychology Quarterly*, 30(2), 188–210. doi: [10.1080/09515070.2016.1169995](https://doi.org/10.1080/09515070.2016.1169995)

- Porter, N. (1995). Supervision of psychotherapists: Integrating anti-racist, feminist, and multicultural perspectives. H. Landrine (Eds.). *Bringing Cultural Diversity to Feminist Psychology: Theory, Research, and Practice*. 163–175. doi:[10.1037/10501-008](https://doi.org/10.1037/10501-008)
- Porter, N., & Vasquez, M. (1997). Covision: Feminist supervision, process, and collaboration. In J. Worell & N. G. Johnson (Eds.), *Shaping the future of feminist psychology: Education, research, and practice* (pp. 155–171). Washington, DC: APA. doi:[10.1037/10245-007](https://doi.org/10.1037/10245-007)
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal Of Educational And Behavioral Statistics*, 4(3), 207–230. doi: [10.3102/10769986004003207](https://doi.org/10.3102/10769986004003207)
- Smith, R. M. (1986). Person Fit in the rasch model. *Educational and Psychological Measurement*, 46(2), 359–372.
- Smith, R. M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement*, 56(3), 403–418.
- Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.
- Stoltenberg, C. D., & McNeill, B. W. (2010). *IDM supervision: An integrated developmental model for supervising counselors & therapists* (3rd ed.). New York, NY: Routledge.
- Szymanski, D. M. (2003). The feminist supervision scale: A rational/theoretical approach. *Psychology of Women Quarterly*, 27, 221–232.
- Szymanski, D. M. (2005). Feminist identity and theories as correlates of feminist supervision practices. *The Counseling Psychologist*, 33, 729–747.
- Walker, A. A., Jennings, J. K., & Engelhard, G. (2018). Using person response functions to investigate areas of person misfit related to item characteristics. *Educational Assessment*, 23(1), 47–68.
- Wolfe, E. W. (2013). A bootstrap approach to evaluating person and item fit to the Rasch model. *Journal of Applied Measurement*, 14(1), 1–9.