Strategies for monitoring mentoring relationship quality to predict early program dropout

Michael D. Lyons | Kelly D. Edwards

INTRODUCTION

Mentoring programs pair nonfamilial adult volunteers with youth to support academic, social, and psychological development (Rhodes & DuBois, 2008). Mentoring programs are one of the most popular, and widely used, prevention interventions deployed in the United States (Garringer et al., 2017). Yet, empirical studies of mentoring programs demonstrate that these programs are associated with small-to-moderate positive effects on a wide range of social, academic, and behavioral outcomes (DuBois et al., 2011; Raposa, Rhodes, et al., 2019). In response to these modest effects, some have suggested that mentoring programs may be able to improve the impact of services by using assessment data to identify, monitor, and evaluate mentors and mentees' (non)response to this service (Cavell et al., 2021; Lyons & McQuillin, 2021). To achieve this goal, programs need to be able to use available data to identify and support mentors and mentees at risk for premature termination. Identifying those at risk for premature termination is important because it has been shown to be associated with iatrogenic effects (Kupersmidt et al., 2017).

In this study, we examined the utility of the Strength of Relationship (SOR) scale (Rhodes et al., 2017) as a potential screening tool for detecting early termination. We draw on data collected from a nationally implemented mentoring program to examine predictors of early termination using a Cox proportional hazards model and then used receiver operating characteristic (ROC) curve analysis to establish empirically derived cut points with associated

Abstract

We examined data from a nationally implemented mentoring program over a 4-year period, to identify demographic and relationship characteristics associated with premature termination. Data were drawn from a sample of 82,224 mentor and mentees. We found matches who reported shared racial or ethnic identities were associated with lower likelihood of premature termination as was mentee's positive feelings of the relationship. We also found that, if data were used as a screening tool, the data were suboptimal for accuracy classifying premature closure with sensitivity and specificity values equal to 0.43 and 0.75. As programs and policymakers consider ways to improve the impact of mentoring programs, these results suggest programs consider the types of data being collected to improve impact of care.

KEYWORDS

measurement, mentoring, youth relationships

Highlights

- Mentors and mentees sharing racial identity had a lower risk of premature termination.
- The quality of the mentoring relationship was, on average, associated with premature termination.
- The quality of an individual mentoring relationship did not predict premature termination.
sensitivity and specificity values. In doing so, the current study provides substantive contributions to research on youth mentoring by identifying correlates of premature termination, as well as provides practical suggestions for mentoring programs to the use of available data in ways that improve service delivery.

Premature termination in mentoring: Why it matters and what can be done?

Youth mentoring refers to a collection of heterogeneous services in which a nonfamilial adult volunteer provides support and guidance to youth (Cavell et al., 2021). Historically, mentoring programs have been based on the logic that providing a child with another supportive adult relationship supports youth social, emotional, and academic outcomes through a close, enduring relationship (Rhodes, 2005). This model of mentoring, sometimes described as a developmental model of mentoring (Karcher et al., 2006) or, more recently, mentoring relationships as an end (Cavell et al., 2021), suggests that programs must optimize training and support in ways that promote long-lasting close relationships between the mentor and the mentee to realize the benefits of mentorship.

As many mentoring programs have approached services from the perspective that a close enduring relationship can be an end in and of itself (Cavell et al., 2021), premature or an unexpected termination of the mentoring relationship is thought to be especially harmful. Empirical studies of these types of mentoring programs have found premature termination can be associated with increased alcohol use, lower perceptions of self-worth, and scholastic competence (Grossman & Rhodes, 2002). In addition, Spencer et al. (2017) found evidence that, when mentoring relationships ended unexpectedly or when mentors abandoned mentees, mentees and their parents were confused, disappointed, and angry about the mentoring experience. It is, therefore, important for mentoring programs to identify, and then support, matches that might be at greater risk for premature termination.

Although many factors (e.g., program practices and logistical barriers) might contribute to premature termination (Spencer et al., 2020), one way that mentoring programs have attempted to mitigate the harm of premature termination and maximize the impact of a mentoring relationship is through matching. The Elements of Effective Practice for Mentoring, a research-informed guide that describes benchmarks for mentoring programs to consider when constructing programs, identifies matching along common characteristics and interests as a core element of practice (MENTOR, 2015). One frequently recommended practice is to pair mentors and mentees together sharing racial, ethnic, or gender identities to facilitate connections across pairs (presumably due to common experiences; Pryce et al., 2013). Albright et al. (2017) explained that

Demographically matched mentors may offer shared life experience that could counter the effects of discrimination and demonstrate possibilities for adolescents’ future selves that are less commonly depicted in their everyday lives or the media. Moreover, an adult with a shared identity could be more relatable. (p. 3).

However, quantitative studies tend to find inconsistent or null associations among match length and demographically matched dyads (DuBois et al., 2011; Raposa, Ben-Eliyahu, et al., 2019). In contrast, qualitative studies have identified racial and gender similarities as both salient and important aspects of a mentoring match (Ellis et al., 2018). One explanation of these mixed effects is that they are difficult to measure quantitatively, because there is limited variability in match characteristics—often resulting in underpowered tests (Albright et al., 2017; Raposa, Rhodes, et al., 2019). Thus, further investigation of possible effects in larger samples is warranted.

To understand how mentoring relationships are developing, programs may measure and track the development of relationship quality over time (e.g., see Spiekermann et al., 2020). Programs are advised to routinely collect data to measure change in the relationship quality over time to support early intervention (Nakkula & Harris, 2013). Many Big Brothers Big Sisters (BBBS) agencies, for example, use the SOR Scale to understand mentor and mentee perceptions of the relationship. This measure was first developed as a self-report scale for mentees to report on their perceptions of the quality of mentoring match (Rhodes et al., 2005). In 2017, this scale was expanded to include mentor perceptions of relationship quality and researchers validated the scales factor structure (Rhodes et al., 2017). The authors also found evidence of the scale’s predictive validity—showing that higher scores on the SOR subscales were associated with longer mentoring relationships. Lyons, Edwards, et al. (2021) also demonstrated that an abbreviated version of the mentor and mentee reports on the SOR demonstrated generally positive reciprocal relations between the mentor and mentee relationship reports over time.

In addition, programs need valid and reliable scales that function similarly across race/ethnicity and gender groups to understand heterogeneity in relationship quality. However, to date, research has not evaluated whether the psychometric properties of the SOR are equivalent across such identity groups. Assessing measurement invariance involves testing the extent to which a scale measures the same latent construct (e.g., relationship quality) with the same degree of accuracy for various groups (Putnick & Bornstein, 2016). If measurement invariance does not hold, then any inferences based on observed group differences are invalid. Consequently, without clear evidence of invariance, programs cannot conclude with certainty whether heterogeneity in SOR scores are the result of true differences in relationship quality or whether those differences...
are simply an artifact of different psychometric properties of the instrument across race/ethnicity and gender groups.

Using available data for person-centered prediction

With greater data available to program staff, there is a call for programs to integrate empirically supported methods to efficiently identify mentors and mentees who may benefit from additional program support (Lyons & McQuillin, 2021). Efficient collection and synthesis of available data could allow mentoring programs to identify matches who are struggling to develop an effective mentoring relationship or make sufficient progress toward goals established by the program or the match. This type of data-informed decision-making mirrors what occurs in other helping relationships (e.g., counseling and psychology), wherein clients are asked to complete assessments to identify areas of need and monitor change over time as well. When integrated into treatment planning protocols, this type of data-informed decision-making has been shown to improve outcomes by allowing helping professionals to identify what may be working (or not working) within the interventions and supports provided (Jenkins et al., 2014). Thus, this may be one strategy for improving impact of mentoring services on youth outcomes.

However, current research on the SOR does not provide information about the specific likelihood of premature termination for a given profile of a specific mentor–mentee match. Instead, prior studies on premature termination have shown relations between individual, ecological, and programmatic correlates of relationship length (Kupersmidt et al., 2017; Raposa, Ben-Eliyahu, et al., 2019). This knowledge provides an understanding of the relation between variables (sometimes referred to as variable-centered analyses) and is useful for understanding associations between constructs of interests and termination (e.g., association between relationship quality and likelihood for premature termination). It does not, however, provide the likelihood for premature termination of a particular mentor–mentee match given their unique data profile (i.e., their SOR scores, match characteristics, etc.).

Is it possible for mentoring programs to use data collected early into a mentoring match, to accurately screen dyads at risk for premature termination? When considering the potential of these data to be used as a screening tool, Glover and Albers (2007) described three key considerations as follows: (1) appropriateness (i.e., alignment between constructs measured and targeted outcomes), (2) usability (i.e., acceptability and feasibility of the measure), and (3) technical adequacy (i.e., reliability and validity of the measure for use as a screening tool). Given prior research on the SOR (Rhodes et al., 2005, 2017) and widespread use of the SOR within BBBS, there appears to be emerging evidence for the SOR for meeting these first two criteria. However, the technical adequacy of the SOR for use as a screening tool has not been evaluated. Specifically, the predictive validity of the SOR for predicting premature termination needs to be established by assessing the sensitivity (i.e., true positives) and specificity (i.e., true negatives) of the measure.

Current study

The current study has three primary aims. First, we examine the psychometric validity of the SOR scales and tested measurement invariance across race and ethnicity, as well as gender. We hypothesized that evidence of measurement invariance would be established. Although a prior study of the psychometric properties of the SOR did not test invariance, we base our hypothesis on prior evidence that the SOR demonstrated construct and predictive validity using a large diverse sample (Rhodes et al., 2017).

Second, we examine predictors of premature termination among mentors and mentees participating in BBBS, the largest mentoring program in the United States. To test this, we used a Cox proportional hazards model in a stepwise manner to estimate the additive predictive effects of (1) demographic and baseline referral characteristics, (2) mentee assessments of the relationship quality, and (3) mentor assessments of the relationship quality. Based on prior work examining demographic characteristics of premature termination, we hypothesize that race, ethnicity, and gender match will have small, significant associations with premature termination such that dyads matched on race and gender will have lower likelihood of premature termination. We also hypothesized that mentee and mentor reports of relationship quality will be modestly associated with premature termination such that higher levels of satisfaction with the mentoring relationship after 3 months will be associated with lower risk for premature termination.

Third, we use the baseline characteristics described above to test the classification accuracy of these predictors. In doing so, we will identify cut points along with sensitivity and specificity values that mentoring programs may use to identify dyads who are at risk for early termination. Although we do not make specific hypotheses about the classification accuracy of these models, we hypothesize that models will demonstrate modest sensitivity and specificity, because the SOR has been shown to be associated with premature termination (Raposa, Ben-Eliyahu, et al., 2019). We expect these results will provide useful information to mentoring programs to make informed choices when selecting baseline assessment tools to deploy among their mentors and mentees.

METHODS

Procedure and participants

Data were collected from BBBS, which operates mentoring services across the United States, from 2014 to 2018. In this study, the analytic sample consisted of
mentors and mentees participating in community-based, BBBS-operated mentoring programs, who also completed the initial demographic questionnaire and one survey assessing the relationship quality (completed at least six months into the mentoring relationship). Out of a total of 362,658 matches from BBBS, 132,670 matches participated in community-based BBBS programs and completed the SOR survey. Of these matches, 82,224 matches provided SOR survey data within 6 months of beginning the mentoring relationship and this sample was used at the analytic sample. A plurality of mentees identified as Black (43.65%) and a majority of mentors identified as White (75.4%; see Table 1 for complete demographics).

TABLE 1 Descriptive statistics

|                | All dyads (N = 82,224) | Terminated early (N = 16,581) | Completed program (N = 65,643) |
|----------------|------------------------|-------------------------------|-------------------------------|
|                | n                      | %                             | n                            | %                             | n                            | %                             |
| **Mentee**     |                        |                               |                              |                               |                              |
| **characteristics** |                        |                               |                              |                               |                              |
| Race           |                        |                               |                              |                               |                              |
| White          | 23,829                 | 34.55                         | 4804                         | 33.97                         | 19,025                       | 34.70                         |
| Black          | 30,106                 | 43.65                         | 6201                         | 43.85                         | 23,905                       | 43.60                         |
| Multiracial    | 11,009                 | 15.96                         | 2290                         | 16.19                         | 8719                         | 15.90                         |
| Other race     | 4027                   | 5.84                          | 847                          | 5.99                          | 3180                         | 5.80                          |
| Hispanic ethnicity | 17,298               | 21.04                         | 3267                         | 19.70                         | 14,031                       | 21.37                         |
| Gender         |                        |                               |                              |                               |                              |
| Female         | 46,784                 | 56.9                          | 9953                         | 60.03                         | 36,831                       | 56.11                         |
| Male           | 35,440                 | 43.1                          | 6628                         | 39.97                         | 28,812                       | 43.89                         |
| Low socio-economic status | 68,309              | 83.08                         | 13,690                       | 82.56                         | 54,619                       | 83.21                         |
| Age (M/SD)     | 11.04                  | 2.41                          | 11.26                        | 2.45                          | 10.98                        | 2.39                          |
| **Mentor**     |                        |                               |                              |                               |                              |
| **characteristics** |                        |                               |                              |                               |                              |
| Race           |                        |                               |                              |                               |                              |
| White          | 57,898                 | 75.43                         | 11,014                       | 71.54                         | 46,884                       | 76.40                         |
| Black          | 9999                   | 13.03                         | 2517                         | 16.35                         | 7482                         | 12.19                         |
| Multiracial    | 3100                   | 4.04                          | 666                          | 4.33                          | 2434                         | 3.97                          |
| Other race     | 5764                   | 7.51                          | 1198                         | 7.78                          | 4566                         | 7.44                          |
| Hispanic ethnicity | 6787                 | 8.25                          | 1471                         | 8.87                          | 5316                         | 8.10                          |
| Gender         |                        |                               |                              |                               |                              |
| Female         | 48,990                 | 59.58                         | 10,449                       | 63.02                         | 38,541                       | 58.71                         |
| Male           | 33,234                 | 40.42                         | 6132                         | 36.98                         | 27,102                       | 41.29                         |
| Age (M/SD)     | 36.84                  | 12.54                         | 35.72                        | 12.29                         | 37.12                        | 12.58                         |
| **Dyad**       |                        |                               |                              |                               |                              |
| **characteristics** |                        |                               |                              |                               |                              |
| Race matched   | 33,554                 | 40.81                         | 7114                         | 42.9                          | 26,440                       | 40.28                         |
| Gender matched | 79,930                 | 97.21                         | 16,069                       | 96.91                         | 63,861                       | 97.29                         |
| Match length in months (M/SD) | 25.1             | 16.51                         | 7.42                         | 2.07                          | 29.56                        | 15.54                         |

Measures

Mentor and mentee demographics

Mentor and mentee demographics were drawn from self-report data collected at the beginning of participation in the program. Participants self-reported data on gender and age. Race and ethnicity were measured as a single self-reported item where participants were asked to indicate one or more racial or ethnic groups to which they self-identified—response options included “American Indian or Native American,” “Asian,” “Black,” “White,” “Other,” “Hispanic,” “two or more races.” Due to small sample sizes of mentors and mentee's endorsing “American Indian or Native American,” “Asian,” “Black,” “White,” “Other,” “Hispanic,” “two or more races” the data was excluded from analyses.
American,” “Asian,” and “two or more races,” we coded these as “Other.” For inferential analyses using the Cox proportional hazards model, two dummy-coded indicators called “race and ethnicity match” and “gender match” were created, to indicate if the mentors and mentees did report the same race/ethnicity or gender (coded as 1), or if their reports were different (coded as 0).

SOR scale

The SOR scale is a measure of mentee-mentor relational satisfaction and attachment. The scale is administered in two versions to obtain reports of relationship quality from mentees and their mentors. The Youth (Y-SOR) version is a 10-item instrument that assesses mentees’ perceptions of positive and negative qualities of the match relationship. Responses are measured using a five-point Likert-type scale, ranging from 1 (never true) to 5 (always true). The Mentor (M-SOR) version includes 14 items that ask mentors to rate their agreement on a five-point scale from 1 (strongly disagree) to 5 (strongly agree). Previous research on the psychometric properties of the SOR found evidence to support a two-factor structure for each informant version of the scale (Rhodes et al., 2017). The two dimensions of the Y-SOR correspond to positive and negative aspects of relationship quality. The M-SOR assesses mentor’s perceptions of affective and logistic dimensions of the mentoring relationship.

Match length and premature termination

Match length was drawn from the program records data and was calculated as the difference in months between the initial match and match closure. Premature termination was calculated as a dichotomous indicator, wherein matches lasting <11 months were coded as one to reflect premature termination and those matches lasting ≥11 months as a zero. Although BBBS sites expect matches that occur in the community to last at least 12 months, we coded premature termination in this way to account for matches that terminated just before the 12-month mark. Of the 82,224 matches included in our analysis, 20% prematurely terminated (Table 1). The average match length was 25.1 months (SD = 16.5) across all matches, but among the subsample of matches that terminated early, the average match length was 7.4 months (SD = 2.1).

Analytic approach

Corresponding to the three primary aims of this study, analyses proceeded in three phases. First, measurement invariance of the SOR was tested across race (Black, White, Multiracial, and other race), ethnicity (Hispanic and non-Hispanic), and gender (female and male) groups. Tests were conducted separately for the Y-SOR and M-SOR versions of the scale. For each set of informants, we evaluated the two-factor measurement model of the scale using Mplus version 8 (Muthén & Muthén, 2017) with mean and variance adjusted diagonally weighted least squares estimation, treating items as ordered categorical. Cluster robust SEs were used to account for the nesting of participants within local mentoring programs. Following recommendations from Wu and Estabrook (2016), and Svetina et al. (2020) for ordinal data, measurement invariance testing involved fitting a series of three increasingly restrictive models that evaluated configurural, threshold, and loading invariance. In the configural model, thresholds and loadings were allowed to vary freely across race, ethnicity, and gender groups. Invariance of the item thresholds was then examined across groups by imposing equality constraints on the thresholds (four for the five response options). Lastly, equality of unstandardized factor loadings was tested with additional constraints imposed on loadings. Model fit was evaluated using the root mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean square residual (SRMR), wherein values of RMSEA ≤ 0.06, CFI ≥ 0.95, and SRMR ≤ 0.08 indicated good fit (Hu & Bentler, 1999).

Nested model comparisons were conducted using $\chi^2$ difference testing. In addition, given that $\chi^2$ tests are sensitive to large sample sizes (Sass, 2011), models were also compared by evaluating changes in fit indices. For ordered categorical indicators, differences in RMSEA < 0.01 and CFI > −0.004 were considered evidence of invariance (Rutkowski & Svetina, 2017). We then conducted a confirmatory factor analysis of the full sample of dyads to obtain factor scores for the four latent variables of relationship quality as measured by the SOR scale.

Second, a stepwise Cox proportional hazards model was estimated, wherein demographic characteristics of the mentor and mentee were included as predictors of premature termination (Step 1) followed by factor scores of mentee-reported relationship quality (Step 2) and mentor-reported relationship quality (Step 3). Time was treated as a continuous variable denoting the number of days to match closure. The conditional probability (i.e., “risk”) of premature termination at time $t$ for a dyad with $x$ covariates is expressed as

$$h(t|x) = h_0(t)\exp(\beta x),$$

where $h_0(t)$ is the baseline hazard function. The estimated hazard ratio (HR), denoted as $\exp(\beta x)$, quantifies the relative risk of early termination between dyads that differ on the set of $x$ covariates. An HR = 1 indicates that all dyads have the same likelihood of early termination and a value below (or above) 1 indicates that the risk is lower (or higher) at any given time, relative to the referent group. In these analyses, the referent group comprised mentees who self-identified as Black males and were matched with non-Black, nonmale mentors at the means of continuous predictor variables (e.g., mean age, mean SOR scores, etc.). Cox regression analysis was conducted in R using the survival package (Therneau, 2021).
Third, the classification accuracy of the available data was assessed by plotting a ROC curve and calculating the area under the curve (AUC). The ROC curve illustrates the proportion of correctly classified dyads (i.e., the true positive rate, or sensitivity) who terminated early against the proportion of incorrect classifications (i.e., the false positive rate, or 1 − specificity) for a given threshold. ROC curves, therefore, provide a useful visualization of the sensitivity and specificity tradeoffs that are made at various threshold levels. As the inflection on these curves maximizes sensitivity and specificity values, it is often used to establish recommended thresholds for screening or clinical purposes. The AUC provides an indicator of overall fit, as it is a summary value of a measure's classification accuracy. Values of AUC range from 0 to 1, where a value of 0.5 indicates the classification accuracy is equal to chance, and a value of 1 indicates perfect predictive accuracy. AUC values of 0.56, 0.64, and 0.71 indicate small, medium, and large effects (Rice & Harris, 2005).

RESULTS
To test whether the measurement properties of the Y-SOR and M-SOR versions were invariant across race, ethnicity, and gender groups (Study aim 1), we compared a series of nested models with increasingly restrictive equality constraints. Fit indices and model comparisons are presented in Table 2. For mentee reports of relationship quality, results of the configural model indicated that the two-factor

TABLE 2 Fit statistics examining measurement invariance by race, ethnicity, and gender

| Model   | RMSEA | CFI  | SRMR | Δχ² | Δdf | ΔRMSEA | ΔCFI |
|---------|-------|------|------|-----|-----|--------|------|
| Y-SOR   |       |      |      |     |     |        |      |
| Race    |       |      |      |     |     |        |      |
| Configural | 0.029 | 0.983 | 0.037 |     |     |        |      |
| Threshold | 0.025 | 0.983 | 0.037 | 66.67 | 60 | −0.004 | −0.001 |
| Loading  | 0.022 | 0.984 | 0.037 | 33.12 | 24 | −0.003 | 0.002 |
| Ethnicity |      |      |      |     |     |        |      |
| Configural | 0.026 | 0.980 | 0.037 |     |     |        |      |
| Threshold | 0.023 | 0.980 | 0.037 | 13.10 | 20 | −0.003 | 0.000 |
| Loading  | 0.021 | 0.981 | 0.037 | 5.73  | 8  | −0.002 | 0.001 |
| Gender  |       |      |      |     |     |        |      |
| Configural | 0.028 | 0.980 | 0.036 |     |     |        |      |
| Threshold | 0.026 | 0.979 | 0.036 | 16.69 | 20 | −0.002 | −0.001 |
| Loading  | 0.024 | 0.981 | 0.036 | 14.63 | 8  | −0.002 | 0.002 |
| M-SOR   |       |      |      |     |     |        |      |
| Race    |       |      |      |     |     |        |      |
| Configural | 0.052 | 0.932 | 0.044 |     |     |        |      |
| Threshold | 0.045 | 0.936 | 0.044 | 129.39* | 84 | −0.007 | 0.004 |
| Loading  | 0.039 | 0.946 | 0.044 | 117.55* | 36 | −0.006 | 0.010 |
| Ethnicity |      |      |      |     |     |        |      |
| Configural | 0.043 | 0.933 | 0.044 |     |     |        |      |
| Threshold | 0.040 | 0.933 | 0.044 | 32.35 | 28 | −0.003 | 0.000 |
| Loading  | 0.036 | 0.940 | 0.044 | 41.08* | 12 | −0.004 | 0.007 |
| Gender  |       |      |      |     |     |        |      |
| Configural | 0.047 | 0.938 | 0.044 |     |     |        |      |
| Threshold | 0.044 | 0.936 | 0.044 | 83.18* | 28 | −0.003 | −0.002 |
| Loading  | 0.041 | 0.939 | 0.044 | 89.22* | 12 | −0.003 | 0.003 |

Note: Comparisons by race included Black, White, Multiracial, and Other race groups. Comparisons by ethnicity included Hispanic and non-Hispanic groups. Abbreviations: CFI, comparative fit index; M-SOR, mentor Strength of Relationship; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; Y-SOR, youth Strength of Relationship.

*p < .05.
solution of the Y-SOR provided good fit across all race (RMSEA = 0.029, CFI = 0.983, SRMR = 0.037), ethnicity (RMSEA = 0.026, CFI = 0.980, SRMR = 0.037), and gender groups (RMSEA = 0.028, CFI = 0.980, SRMR = 0.036). When item thresholds were constrained to be equal across groups, \( \chi^2 \) difference tests revealed that imposing the constraints was tenable for race, \( \Delta \chi^2 (60) = 66.67, p = .26 \); ethnicity, \( \Delta \chi^2 (20) = 13.10, p = .87 \); and gender, \( \Delta \chi^2 (20) = 16.69, p = .67 \). In addition, changes in fit statistics between the configural and threshold models were within the recommended cutoffs for race (\( \Delta \text{RMSEA} = -0.004, \Delta \text{CFI} = -0.001 \)), ethnicity (\( \Delta \text{RMSEA} = -0.003, \Delta \text{CFI} = 0.000 \)), and gender (\( \Delta \text{RMSEA} = -0.002, \Delta \text{CFI} = -0.001 \)), providing further evidence of threshold invariance for the Y-SOR. In the third model, additional equality constraints were imposed on factor loadings. Chi-square difference tests supported invariance of factor loadings for race, \( \Delta \chi^2 (24) = 33.12, p = .10 \); ethnicity, \( \Delta \chi^2 (8) = 5.73, p = .68 \); and gender, \( \Delta \chi^2 (8) = 14.63, p = .07 \), and nominal changes in fit indices were observed.

For mentor-reported relationship quality, the configural model demonstrated adequate fit across categories of race (RMSEA = 0.052, CFI = 0.933, SRMR = 0.044), ethnicity (RMSEA = 0.043, CFI = 0.933, SRMR = 0.044), and gender (RMSEA = 0.047, CFI = 0.938, SRMR = 0.044). Nearly all \( \chi^2 \) difference tests of the M-SOR models were statistically significant; however, changes in fit indices revealed that model fit generally improved as additional constraints were imposed. Notably, threshold invariance was supported for race (\( \Delta \text{RMSEA} = 0.007, \Delta \text{CFI} = 0.004 \)), ethnicity (\( \Delta \text{RMSEA} = -0.003, \Delta \text{CFI} = 0.000 \)), and gender (\( \Delta \text{RMSEA} = -0.003, \Delta \text{CFI} = -0.002 \)), as was loading invariance across race (\( \Delta \text{RMSEA} = -0.006, \Delta \text{CFI} = 0.010 \)), ethnicity (\( \Delta \text{RMSEA} = -0.004, \Delta \text{CFI} = 0.007 \)), and gender (\( \Delta \text{RMSEA} = -0.003, \Delta \text{CFI} = 0.003 \)).

Results illustrating the relations between mentor and mentee characteristics and premature termination analyzed using Cox proportional hazard analysis (Study aim 2) are reported in Table 3. In Model 1, mentee and mentor age, mentee race/ethnicity, and dyad race/ethnicity and gender match were significant predictors of early termination. Whereas a 1-year increase in mentee age was associated with a 5% increased risk of premature termination (HR = 1.05, p < .001), a 1-year increase in mentor age was associated with a 1% reduced risk (HR = 0.99, p < .001). Comparisons between mentee race/ethnicity groups indicated that White and multiracial mentees had a 21% risk increase (HR = 1.21, p < .001) and 12% (HR = 1.12, p = .002) higher risk of early termination, respectively, when compared to Black mentees. Although gender match was associated with a 25% reduced risk of terminating early (HR = 0.75, p < .001), race/ethnicity match was associated with a 34% increased risk (HR = 1.34, p < .001) after controlling for all other model covariates. Tests of interactions between race/ethnicity match and mentee race/ethnicity categories revealed that, in comparison to Black mentees matched with non-Black mentors, the relative risk of premature termination was 34% lower for White mentees matched with White mentors (HR = 0.66, p < .001), 18% lower for Hispanic mentees with Hispanic mentors (HR = 0.82, p = .005), and 24% lower for multiracial mentees with multiracial mentors (HR = 0.76, p = .006).

Model 2 tested whether mentee perceptions of match relationship quality predicted early termination after accounting for baseline demographic characteristics. We found that scores on the “Positive” subscale of the Y-SOR were predictive of premature termination; however, scores on the “Negative” subscale were not. Controlling for baseline covariates, a 1 SD increase on the Y-SOR-positive subscale was associated with a 26% reduced risk of terminating early (HR = 0.74, p < .001). Mentor-reported relationship quality was included as a predictor in Model 3. Results showed that both M-SOR subscales significantly predicted early termination. A 1 SD increase on the M-SOR affective subscale lowered the risk of premature termination by 24% (HR = 0.76, p < .001). Likewise, a 1 SD increase on the M-SOR logistic subscale lowered the risk by 9% (HR = 0.91, p < .001).

ROC analyses were conducted to evaluate how well the SOR scale (combined Y-SOR and M-SOR scales) differentiated matches that terminated early and those that did not (Study aim 3). The relationship between the true-positive rate and false-positive rate is illustrated with the ROC curve in Figure 1. The overall accuracy of the SOR scale to predict premature terminations was modest (AUC = 0.63, 95% confidence interval = 0.63, 0.64). Sensitivity and specificity were calculated for various threshold values of the SOR scale. As shown in Table 4, sensitivity values ranged from 0.05 to 0.95 and specificity values ranged from 0.15 to 0.96. A scale score of 4.25 was identified as the optimal threshold value with sensitivity and specificity equal to 0.43 and 0.75, respectively. Here, this threshold value can be interpreted as the SOR value below which programs may identify students as likely to prematurely terminate after controlling for demographic characteristics.

**DISCUSSION**

Youth mentoring programs are popular prevention services that are often thought to produce positive effects through a close, long-lasting relationship with a non-familial adult (Rhodes, 2005). This means that unexpected and premature termination of the mentoring relationship can be especially harmful for youth (Spencer et al., 2017). To mitigate risks of premature termination, programs are moving toward models of service in which mentoring activities are guided by research-based practices and supported by data (Lyons & McQuillin, 2021). BBBS, for example, routinely collects data from mentors and mentees, and employs staff to support and advise participants in effective strategies for mentoring. Although the data BBBS collects on the match characteristics and relationship strength has been shown to correlate with important outcomes (e.g., match length and positive reciprocal
As mentoring programs move to use available data to make informed, research-based decisions about the types of services offered, the results of the study suggested that existing data collection efforts within BBBS provide a promising foundation upon which to begin to identify those at risk for prematurely terminating the service. As data were collected from a large national sample of mentors and mentees participating in the program (N = 82,224), results presented provide new evidence about the ways in which the assessment data do (and do not) provide a signal for those who may be at risk for termination. Three major findings related to the construct validity, predictive validity, and classification accuracy of the available data are described below.

First, we conducted analyses to establish the construct validity of the SOR as an indicator of premature termination. Building on prior work establishing strong psychometric validity of the tool (Rhodes et al., 2017), we found additional support for measurement invariance across race and ethnicity given the factor structure of the measure. This means that items and latent constructs measured by the SOR (e.g., mentee's positive feelings about their mentor) appear to function in the same way (i.e., similar factor loadings) for different racial and ethnic groups participating in BBBS in the United States. These results provide additional evidence that statistically or

| TABLE 3 Stepwise logistic regressions predicting premature terminations |
|------------------|---|---|---|---|---|---|---|---|
|                | Model 1 |         | Model 2 |         | Model 3 |         |
|                | HR    | SE    | 95% CI | HR    | SE    | 95% CI | HR    | SE    | 95% CI |
| Mentee age     | 1.05* | 0.003 | 1.04, 1.06 | 1.03* | 0.005 | 1.02, 1.04 | 1.03* | 0.005 | 1.02, 1.04 |
| Mentee SES     | 0.97  | 0.021 | 0.93, 1.02 | 0.98  | 0.027 | 0.94, 1.04 | 0.98  | 0.027 | 0.93, 1.03 |
| Mentor age     | 0.99* | 0.001 | 0.99, 0.99 | 0.99* | 0.001 | 0.99, 0.99 | 0.99* | 0.001 | 0.99, 0.99 |
| Mentee race/ethnicity (reference group: Black) |
| White          | 1.21* | 0.044 | 1.10, 1.33 | 1.20* | 0.048 | 1.09, 1.31 | 1.23* | 0.048 | 1.12, 1.35 |
| Hispanic       | 0.93  | 0.028 | 0.82, 1.05 | 0.93  | 0.062 | 0.83, 1.05 | 0.96  | 0.067 | 0.84, 1.10 |
| Multiracial    | 1.12*** | 0.026 | 1.04, 1.20 | 1.12* | 0.036 | 1.05, 1.20 | 1.16* | 0.035 | 1.08, 1.24 |
| Other race     | 1.08  | 0.042 | 0.95, 1.23 | 1.06  | 0.061 | 0.94, 1.20 | 1.10  | 0.063 | 0.97, 1.25 |
| Gender (reference group: Male) |
| Female         | 0.72  | 0.356 | 0.37, 1.40 | 0.66  | 0.372 | 0.32, 1.38 | 0.60  | 0.386 | 0.28, 1.27 |
| Match characteristics |
| Race/ethnicity matched | 1.34* | 0.027 | 1.23, 1.46 | 1.26* | 0.042 | 1.16, 1.37 | 1.30* | 0.042 | 1.20, 1.42 |
| Gender matched | 0.75* | 0.047 | 0.67, 0.84 | 0.75* | 0.056 | 0.67, 0.84 | 0.75* | 0.056 | 0.67, 0.84 |
| Race/ethnicity match interactions |
| Matched × White | 0.66* | 0.052 | 0.58, 0.74 | 0.72* | 0.062 | 0.64, 0.82 | 0.70* | 0.062 | 0.62, 0.79 |
| Matched × Hispanic | 0.82*** | 0.054 | 0.71, 0.94 | 0.85**** | 0.066 | 0.75, 0.97 | 0.83*** | 0.068 | 0.73, 0.95 |
| Matched × Multiracial | 0.76*** | 0.100 | 0.62, 0.92 | 0.83 | 0.101 | 0.68, 1.01 | 0.81*** | 0.101 | 0.67, 0.99 |
| Matched × Other race/ethnicity | 0.91 | 0.088 | 0.74, 1.12 | 0.90 | 0.095 | 0.75, 1.08 | 0.85 | 0.103 | 0.70, 1.04 |
| Gender matched × Female | 1.60 | 0.357 | 0.82, 3.12 | 1.79 | 0.375 | 0.86, 3.74 | 1.96 | 0.390 | 0.91, 4.21 |
| Mentee SOR |
| Negative subscale | 1.00 | 0.001 | 0.98, 1.02 | 1.01 | 0.010 | 0.98, 1.03 |
| Positive subscale | 0.74* | 0.002 | 0.72, 0.75 | 0.84* | 0.013 | 0.82, 0.86 |
| Mentor SOR |
| Affective subscale | 0.76* | 0.012 | 0.75, 0.78 |
| Logistic subscale | 0.91* | 0.012 | 0.88, 0.93 |

Note: Race/ethnicity and gender match coded as: 0, mentors and mentees did not report same race/ethnicity or gender; 1, mentors and mentees reported the same race/ethnicity or gender.
Abbreviations: CI, confidence Interval; HR, hazard ratio; SOR, Strength of Relationship.
*p < .001; **p < .05; ***p < .01.
clinically significant differences in self-reported SOR scores may be attributed to factors other than the psychometric properties of the scale. As mentoring programs move toward models that more explicitly address and support racially and ethnically diverse mentors and mentees, results from the SOR scales may be one measure that captures substantive differences in mentors and mentees experiences with their mentoring relationship.

Second, we found evidence that demographic and select SOR subscales were significantly associated with premature termination. Results suggested that mentors and mentees who self-reported a shared racial or ethnic background was associated with lower likelihood of premature termination as compared to Black mentees matched with non-Black mentors—of which, the majority, (83%) of non-Black mentors identified as White. This effect was strongest among Black mentors and mentees who were, on average, estimated to have a 30% lower likelihood of premature termination. This finding may suggest that matching mentors and mentees with shared racial or ethnic background could be one strategy for reducing the likelihood of premature termination. As Albright et al. (2017) suggested, matching may promote positive mentoring outcomes, because it may allow mentors additional opportunities to promote positive racial identity and support youth who may be experiencing racialized stressors.

We also observed small-to-moderate associations between SOR subscales and premature termination. Sub-scales that assessed positive attitudes about the mentoring relationship appeared to be more salient predictors of premature termination as compared to other subscales of the SOR. These findings are consistent with prior theory on mentoring relationships, suggesting that mentors and mentees self-assessed attitudes about the quality of the relationship are associated with duration of the match (Grossman & Rhodes, 2002). Thus, based on findings from this study and prior research, it appears that indicators of relationship quality may be associated with reductions in the overall likelihood for premature termination. These results are consistent with the prior theory that average increases in relationship quality are associated with a reduced risk of premature termination.

Third, and perhaps counterintuitively, we only found modest evidence for the use of available data as a means for accurately classifying those at risk for premature termination. Specifically, thresholds identified using ROC analyses found sensitivity values ranging from 0.05 to 0.95 and specificity values ranging from 0.15 to 0.96 with optimal threshold value equal to 4.25 with sensitivity and specificity equal to 0.43 and 0.75, respectively. As compared to screening tools used in other applications (e.g., mental health screeners), these values fall below typically acceptable sensitivity values. This means that, if used as a screening tool, SOR data may be unable to make sufficiently precise predictions as to which dyads are likely to prematurely termination from the program. In other words, programs may use SOR data to draw conclusions about average associations between relationship quality and premature termination; however, these data are not sufficient to accurately identify specific dyads at risk for premature termination.

For programs and researchers wishing to use available data as a screening tool to inform actionable steps to support youth outcomes, this finding may be disappointing. Nevertheless, these data are useful for understanding program-level average effects that could be used to improve mentor and mentee participation in programming. For example, findings suggest that, on average, matching
mentees by race and ethnicity may substantially reduce the likelihood for premature termination. One potential implication of this finding is to continue to diversify the available pool of mentors to enable more widespread matching as a method for reducing the likelihood of premature termination. In addition, programs may also pursue targeted training and support for mentors (primarily White) matched with historically minoritized youth to ensure mentoring practices are culturally responsive and help facilitate mentor and mentee engagement while limiting premature termination.

**LIMITATIONS AND FUTURE DIRECTIONS**

Despite various strengths of the current study, there are at least three limitations and areas for future research. First, predictions for premature termination from the mentoring program were based on demographic information and SOR data collected in the first six months from mentors and mentees participating in the program. Although this choice was made to include the largest number of participants, it is possible that more routine collection of SOR indicators would provide more robust predictions of premature termination. Future studies may wish to examine the classification accuracy of the SOR as a function of the frequency of data collection. In addition, some studies have suggested that when relationship quality is considered in the context of other mentoring activities (e.g., setting goals and providing feedback), stronger effects of mentoring are observed (Lyons et al., 2019); thus, future studies may wish to jointly assess relationship strength as well as progress toward mentee-directed goals as indicators of premature termination.

Second, the results suggested premature termination was moderately reduced when mentors were matched by race or ethnicity (as compared to Black mentees matched with non-Black mentors). This finding is consistent with some, but not all, prior quantitative studies examining associations between race and ethnicity match and youth outcomes (DuBois et al., 2011). Given these mixed findings, it may be useful for additional studies to examine the hypothesized mechanisms (e.g., affirming racial identity) by which race/ethnicity matching is thought to produce these positive effects (Albright et al., 2017). Understanding the mechanisms by which race or ethnicity match influences termination would provide a more precise explanation for why race and ethnicity matching appears to be a protective feature to guard against premature termination. For example, although study results found significant reductions in risk of premature termination among mentors and mentees identifying and multiracial, it is not obvious why this may be a relation observed given that this may reflect a heterogeneous subgroup of mentor–mentee matches. It is important to emphasize that race and ethnicity are socially constructed, meaning that differences observed across identities should be attributed to social, cultural, or political factors that shape how racial differences might influence adolescent development. Garcia-Coll et al. (1996) outlined one theoretical model that might be useful in understanding how socio-cultural factors, such as racism, might influence the development of mentoring relationships. In addition, other factors unrelated to the racial or ethnic identity of the mentor and mentee might be contributing factors that influence the likelihood of premature termination. Spencer et al. (2020) wrote, for example, that “it takes a village to break up a match” referring to the variety of community, family, and programmatic features that contribute to the duration of mentoring matches. Program training and ongoing match support, for example, is one programmatic protective factor identified as being related to reducing risk of premature termination (McQuilllin & Lyons, 2021).

Third, the results presented should not be interpreted as causal—that is, in this study, poor relationship quality or demographic characteristics were not found to be the reason for premature termination. Given modest associations between the measured variables and termination, other unmeasured variables may be contributing to the observed relations. Although establishing a causal relationship is not always necessary (Grosz et al., 2020), understanding causal factors by conducting randomized controlled trials or creating a comparison condition using statistical or other means (e.g., propensity score matching and sequential multiple assignment randomized trials) can be useful as a means for understanding how specific mentoring practices may mitigate this risk. Future studies may wish to test how specific programmatic activities (selected using best available data) can mitigate risk of premature termination.

**CONCLUSIONS**

Mentoring programs are among the most popular and widely implemented types of prevention programs. In response to calls to shift mentoring services in ways that are aligned with evidence-based practices, understanding how available data routinely collected can be used to inform programmatic decisions is an important first step (McQuilllin et al., 2019). Although data-based decision-making is an important element identified in many helping professions (Glover & Albers, 2007), it is essential among paraprofessional and volunteer services, like mentoring, because these services are increasingly identified to expand access to support services to historically underserved populations (Cavell et al., 2021). Using and expanding mentoring programs’ capacity to collect and synthesize available data in ways that directly inform services may be one approach for achieving this goal.

**CONFLICT OF INTERESTS**

The authors declare that they have no conflict of interests.
SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher’s website.