American Psychological Association

Psychological Assessment

© 2018 American Psychological Association
1040-3590/19/$12.00

2019, Vol. 31, No. 3, 352–364
http://dx.doi.org/10.1037/pas0000672

Efficiently Assessing Firm, Fair, and Caring Relationships: Short Form of the Dual Role Relationship Inventory

Perman Gochyyev and Jennifer L. Skeem
University of California, Berkeley

Many people involved in the justice system and people with serious mental illness are required to participate in psychosocial treatment, whether they want it or not. With these clients, case managers, probation officers, and other providers are tasked with both promoting client recovery (a helping, therapeutic role) and protecting community safety (a controlling, surveillance role). The 30-item revised Dual-Role Relationship Inventory (DRI-R) assesses the quality of provider–client relationships in mandated treatment—and DRI-R based research indicates that firm, fair, and caring relationships (authoritative, not authoritarian) predict better client outcomes. In this study, we developed and validated a short form of the DRI-R—the 9-item DRI-SF—by applying multidimensional item response theory methods to four data sets ($N = 815$). We simultaneously refined the measure by selecting items that cleanly assessed relationship features (i.e., minimized construct-irrelevant variance from provider traits) and performed similarly across client groups (juveniles and adults; with-and-without mental illness). DRI-SF total scores strongly predict DRI-R total scores ($r = .97$). The DRI-SF fully represented the DRI-R's range of item difficulties, produced the same three-factor structure, predicted theoretically relevant external covariates as strongly (i.e., groups known to differ in relationship quality, relationship satisfaction ratings, future arrests)—without item bias by sex or race. Moreover, the favorable psychometric properties of the DRI-SF were replicated in a new sample and shown to generalize across provider groups (from probation officers to treatment providers). This newly developed DRI-SF applies to a range of provider–client relationships in mandated treatment—and will benefit practitioners and researchers with ease of administration.

Public Significance Statement
When treatment is mandated, high-quality relationships between clients and their providers promote positive outcomes. In this study, we used a data-informed approach to shorten a well-validated measure of firm, fair, and caring client–provider relationships. This clinically feasible measure can be readily applied with a broad range of clients in busy mental health, justice, and research settings.

Keywords: therapeutic alliance, dual-role relationship, mandated treatment, item response theory, differential item functioning

Supplemental materials: http://dx.doi.org/10.1037/pas0000672.supp

People who are involved in the justice system and people with serious mental illness are often formally or informally required to participate in psychosocial treatment (Monahan et al., 2005). Clients involved in the justice system can be legally mandated to take part in substance abuse, correctional, and mental health services. Similarly, case-management services can be "assertively" provided to clients with serious mental illness in the community. In these contexts, classic notions of the therapeutic alliance—which relatively strongly predicts client outcomes (e.g., Horvath, Del Re, Flückiger, & Symonds, 2011)—rarely capture the quality of the relationship between a client and service provider. When clients are required to take part in treatment, service providers have dual roles: They not only care for, but also have control over, the client (Skeem, Louden, Polaschek, & Camp, 2007). Case managers, probation officers, and other providers are tasked both with promoting client recovery (a helping, therapeutic role) and protecting community safety (a surveillance role; Trotter, 1999).

Theoretically and empirically, "firm, fair and caring" dual role relationships promote better outcomes than relationships that are strictly therapeutic or strictly authoritarian (Kennealy, Skeem, Manchak, & Eno Louden, 2012; Skeem & Manchak, 2008). The basis for this statement rests partly on the 30-item Dual-Role Relationship Inventory (DRI-R; Skeem et al., 2007), a well-validated measure of the quality of provider–client relationships in

This article was published Online First November 5, 2018.

Perman Gochyyev, Graduate School of Education, University of California, Berkeley; Jennifer L. Skeem, School of Social Welfare, Goldman School of Public Policy, University of California, Berkeley.

Correspondence concerning this article should be addressed to Perman Gochyyev, Graduate School of Education, University of California, Berkeley, 2121 Berkeley Way #4205, Berkeley, CA 94720. E-mail: perman@berkeley.edu

mandated treatment that was originally developed in the context of probation and has since been more widely applied. Although authoritative (rather than authoritarian) relationships have long been regarded as a key component of effective practice with justice-involved people (Dowden & Andrews, 2004), the DRI-R appears to be the only well-validated operationalization of this construct.

In the present study, we use sophisticated analytic techniques with data from four different studies to develop and validate a short (nine-item) form of the DRI-R—with two general purposes in mind. The first purpose is to offer an efficient measure of relationship quality in mandated treatment that can be widely used in research and practice. The second is to use the opportunity to refine the scale by ensuring that the short form cleanly measures relationship features (rather than construct-irrelevant provider traits), generalizes across client groups (adult and youth; with- and without mental illness), and generalizes across professional groups (probation officers and treatment providers). We also test for validity across client sex and race.

Need for a Feasible Measure of Dual Role Relationship Quality

Dual Role Relationships Are Common

As observed by Howgego, Yellowlees, Owen, Meldrum, and Dark (2003), conventional measures of the therapeutic alliance poorly fit clients who "do not voluntarily seek help and enter a relationship motivated to engage" (p. 180). Of the 4.7 million people on probation or parole in the United States (Kaeble & Glaze, 2016), a substantial proportion are legally mandated to participate in substance abuse, correctional, or mental health services. Of the 9.8 million people with serious mental illness in the United States (Center for Behavioral Health Statistics and Quality, 2016), a subset doubt the reality of their illness or the value of treatment offered—and are subject to a variety of formal and informal pressures to comply (Burns, 2016).

Formal treatment mandates come in many forms, including involuntary outpatient commitment and special conditions of probation, both of which involve judicial orders to adhere to a community treatment plan. Based on a sample of 1,000 outpatients drawn from public community outpatient settings, Monahan et al. (2005) found that nearly half (44%–66%) had experienced at least one of four types of formal mandates to participate in treatment (for U.K. rates, see Burns et al., 2011).

Clients are also subject to informal pressure from providers to adhere to treatment. For example, case management models for people with serious mental illness often prioritize engagement. The rationale behind the widely disseminated Assertive Community Treatment (ACT) team model "has always been to engage patients in treatment that they fundamentally do not want" (Burns, 2016, p. 15). Clinicians use a range of techniques to "encourage, persuade, and cajole . . . patients to comply with their prescribed treatment" (Burns, 2016; see Angell & Mahoney, 2007; Angell, Mahoney, & Martinez, 2006). Based on a sample of 1,564 veterans treated by ACT teams, Neale and Rosenheck (2000) found that case managers routinely used strong verbal guidance and often used money management to control behavior. Less often, they used contingent withholding of help, hospitalization, and appeals to external authorities.

Conventional measures of the alliance do not capture the control inherent in these relationships. Specifically, a wealth of psychotherapy research suggests that the client-provider relationship reflects an accumulation of interpersonal interactions over time that vary in their degree of (a) affiliation or connectedness (ranging from hostile to friendly) and (b) control or influence (ranging from controlling to autonomy granting on the part of the provider or from submissive to autonomy-taking on the part of the client; see Benjamin, Rothweiler, & Critchfield, 2006; Henry, Schacht, & Strupp, 1990; Kiesler, 1983). Based on an assessment of 125 provider–client relationships, Manchak, Skeem, and Rook (2014) found that relationships in mandated or assertive treatment involve greater control than those in voluntary treatment—but can remain largely affiliative. In keeping with well-established circumplex models of interpersonal behavior, this finding helps illustrate that control and affiliation are independent dimensions. That is, control can be applied in a manner that is hostile, neutral, or affiliative (Benjamin et al., 2006).

Dual Role Relationship Quality Predicts Outcomes

A large body of research attests to the power of the therapeutic alliance in predicting outcomes that range from symptom improvement to reduced violence (for a review, see Lambert, 2013). In fact, meta-analyses of psychotherapy research suggest the quality of the therapeutic alliance out-predicts other controllable predictors of client outcomes, including the type of techniques applied (e.g., Horvath et al., 2011; Martin, Garske, & Davis, 2000).

A smaller body of research on mandated treatment indicates that high quality dual role relationships also predict positive outcomes, although most focus has been on criminal recidivism. For example, based on a meta-analysis of 273 effect sizes, Dowden and Andrews (2004) found that correctional agencies staffed by warm, empathic, respectful and nonblaming officers were relatively effective in reducing recidivism. More recently, firm, fair, and caring officer–client relationships—as operationalized by the DRI-R (Skeem et al., 2007)—have been shown to protect against recidivism among probationers with mental illness (Skeem et al., 2007; Skeem & Manchak, 2010), parolees without mental illness (Kennealy et al., 2012), and juvenile offenders (Vidal, Oudekerk, Reppucci, & Woolard, 2013). The DRI-R predicts such outcomes more strongly than measures of the traditional therapeutic alliance (Skeem et al., 2007). It is our hope that dual role relationship quality will be more routinely studied to determine its impact on a wider variety of outcomes.

Dissemination of An Instrument Is Influenced by Its Length

Intensive study of the therapeutic alliance has been facilitated by longstanding, high-quality self-report operationalizations of the construct that has been made more efficient over time. For example, the well-validated 36-item Working Alliance Inventory (Horvath & Greenberg, 1989) has been condensed into a 12-item short form (Hatcher & Gillaspy, 2006) that is used even more widely than the original.

In both practice and research, short forms require less time to complete, are less burdensome to clients (who may not be enthusiastic), and are more appropriate for repeated measurements over time. In front-line clinical and justice settings where dual role relationships are most common, resources are scarce and formal assessment tools are rarely used. This underscores the need for an efficient measure. As shown next, we now have the data and analytic tools needed to develop a short form of the 30-item DRI-R (Skeem et al., 2007) that may have better psychometric properties than the parent scale. The development of short forms offers opportunities for scale refinement.

DRI-R

The DRI-R was created through a multistage process that involved a qualitative study ($n = 52$) followed by two quantitative studies ($n = 90$; $n = 322$) and included item development, item refinement, and measure validation. Details are provided in Skeem et al. (2007) and briefly summarized here.

First, we identified contours of relationship quality in mandated treatment by conducting a multisite focus group study with people with mental illness and their supervising probation officers (Skeem, Encandela, & Louden, 2003). Participants described two general types of officer–client relationships: authoritarian relationships characterized by many demands, little flexibility, and belittling use of control (which often led to client stress, avoidance or reactance); and (b) authoritative relationships characterized by warmth and social control that was used in a manner perceived as fair, respectful, and motivated by caring (which often led to client trust, disclosure, and compliance).

Second, we used these qualitative results—along with leading conceptualizations and measures of the therapeutic alliance (relevant to the caring role) and procedural justice (relevant to the controlling role)—to draft a pool of potential DRI items. We then administered the DRI and external validation measures in a multimethod study, which we conducted in a prototypic dual role relationship context—specialty mental health probation. Compared to traditional probation, specialty probation is distinguished by small caseloads (<50 individuals) composed solely of people with mental illness (rather than heterogeneous caseloads with >100 individuals); sustained officer training in mental health; and heavy officer involvement in clients' treatment (Skeem, Emke-Francis, & Louden, 2006). In specialty agencies, officers balance "control" (surveillance) with "care" (rehabilitation) and stress engagement in psychiatric and other services. Third and finally, we administered the 30-item DRI-R to probationers with mental illness supervised in either traditional- or specialty-probation to replicate and extend initial findings.

Results of these studies indicated that the DRI-R is internally consistent; relates in a theoretically coherent pattern to both observer ratings of within-session behavior (e.g., confrontation, resistance) and validated measures of the therapeutic alliance, relationship satisfaction, and treatment motivation; distinguishes between specialty and traditional probation groups known to differ in relationship quality; and predicts compliance with the rules (Skeem et al., 2007; see also Kennealy et al., 2012, and Vidal et al., 2013, for additional evidence of reliability, convergent validity, and predictive utility). Confirmatory factor analyses indicated that the DRI-R assesses three dimensions: (a) caring and fairness, which includes warmth; clear communication about rules, roles, and limits; interest in the client's "voice" and perspective; and respect; (b) trust, which taps the client's feeling of safety to disclose problems and the providers' trust in the client; and (c) toughness, which captures an authoritarian style and is weighted inversely in DRI-R total scores (Skeem et al., 2007). High scores on the DRI-R signify authoritative (not authoritarian) professional-client relationships in mandated treatment—firm, fair, and caring.

Present Study Aims

In the present study, we carefully analyze data from four different studies to develop and validate a short form of the DRI-R that can be used widely in research and practice. We aimed to reduce the number of items on the DRI-R from 30 to 9 because nine items represent a version that is as short as possible, but no shorter than required to achieve desirable psychometric properties. Specifically, a nine-item short form both (a) provides for equal representation of each factor with three items, without compromising reliability at the factor level (as would be the case with fewer items; e.g., Velicer & Fava, 1998); and (b) reduces the scale to less than one third of its original length, making survey administration considerably more efficient.

We also used this opportunity to refine and extend the DRI-R. First, we prioritize items for the short form that measure the quality of a specific client-provider relationship as cleanly as possible, without undue influence by the provider's personal characteristics. As shown later, DRI-R scores tend to weakly cluster by provider, suggesting that a non-negligible part of the variance is attributable to provider characteristics (e.g., traits, styles, attitudes) that may be measured separately. Second, we prioritize items for the short form that apply equally well across client groups (e.g., with- or without mental illness; adult or juvenile). Although the DRI-R was developed with justice-involved people with mental illness, it has since been applied to other client populations, allowing the short form to be approached with an eye toward generalizability. Third, we compare the short form's performance across professional groups. The DRI-R was developed with specialty probation officers and has mainly been used to assess officer–client relationships—but has since been used to assess relationships between mandated treatment providers and clients. This permits a test of generalizability across service provider types.

In sum, this study's first aim is to develop a nine-item short form of the DRI-R that both (a) measures relationship quality cleanly, minimizing construct-irrelevant variance contributed by officer traits; and (b) performs similarly across client groups (i.e., adult/juvenile; with/without mental illness). The study's second aim is to validate the short form by determining whether it distinguishes between known groups, correlates with relationship satisfaction, and predicts recidivism as strongly as the full DRI-R—and by testing whether its psychometric properties can be replicated in—and extended to—a new sample that focuses on provider-client (not officer-client) relationships. We also test the short form for validity across client variation in sex and race.

Method

These aims were addressed chiefly by using a variety of item response theory (IRT) techniques to analyze data on the DRI-R and other relevant measures. Next, we summarize the data sets we used and then describe the IRT and external validation analyses.

Data Sets

Four data sets were used to construct and validate the short form of the DRI-R. Combined, these data sets represent 815 offenders—both adults and juveniles, both with- and without mental illness—and include ratings of relationships with both supervising officers ($n = 690$) and treatment providers ($n = 125$). We highlight each dataset below (details are available in the references cited), before describing how the data were used together to achieve the study aims.

1. The "main" dataset (Skeem, Manchak, & Montoya, 2017; see also Manchak, Skeem, Kennealy, & Louden, 2014) includes 359 adult probationers with mental illness, 176 of whom were supervised by traditional officers ($n = 82$), and 183 of whom were supervised by specialty mental health officers ($n = 15$). Probationers' average age was 36.9 ($SD = 10.6$). Of probationers, 57% were male and 50% were African American (38% non-Hispanic Caucasian; 9% Hispanic; and 3% other). Beyond the probationer-rated version of the DRI-R at baseline, we also used the following measures from this dataset: (a) probationers' baseline ratings of satisfaction in their relationship with their probation officers (rated on a 5-point Likert scale), and b) time to recidivism, defined as whether or not any arrest for a new offense occurred within a minimum 2-year follow-up period and the date of any arrest or maximum date of follow-up if there was no arrest (which has mean of 348 days and 1,117 days for probationers with any arrest and no arrest, respectively).

2. The "juvenile" dataset (Vidal & Woolard, 2017) includes 110 juvenile probationers with an average age of 16.4 ($SD = 1.4$). Of juveniles, 77% were male, 60% were African American (11% Hispanic, 7% non-Hispanic Caucasian; 11% other), and 51% had been referred to the juvenile justice system three or more times.

3. The "mixed adult" dataset (Skeem, Winter, Kennealy, Louden, & Tatar, 2014) includes 221 adult parolees, 112 of whom had mental illness and a demographically matched sample of 110 of whom did not have mental illness. Parolees' average age was 39 ($SD = 9$); 88% were male and 71% were African American (11% Hispanic, 7% non-Hispanic Caucasian; 11% other).

4. The "provider" dataset (Manchak, Skeem, & Rook, 2014) include 125 mental health court participants who were required to participate in treatment. Unlike other participants (who used the DRI-R to rate their relationships with supervising officers), mental health court participants used the DRI-R to rate the quality of their dual role relationship with their primary treatment provider. Participants' average age was 37 ($SD = 11.4$); 54% were women and 67% were non-Hispanic White (16% Hispanic, 10% African American).

To develop the DRI-R short form (i.e., address Aim 1), we merged the first three data sets (i.e., the main, juvenile, and mixed adult data sets) and mostly conducted item response modeling on this "combined" dataset, which includes 690 offenders. To externally validate the DRI-R short form, we used the main dataset and the provider dataset, as detailed below. No institution review board approval was required because all four data sets were deidentified.

Analyses

Item response modeling to develop a nine-item short form. We used IRT methods to analyze the items of the DRI-R and its nine-item short form. Item analysis is a set of qualitative and quantitative approaches that can be used to inform the selection (and omission) of items from an instrument (Gochyyev & Sabers, 2009). IRT methods have advantages over classical test theory approaches (Cooke, Michie, Hart, & Hare, 1999; Embretson & Reise, 2000)—and may be viewed as confirmatory factor analysis for categorical variables (Brown, 2006).

Within the family of IRT models, we primarily use the Rasch model—a one-parameter logistic model or 1PL. We do so because, relative to alternatives like a two-parameter logistic IRT model (or 2PL), the Rasch model fit the data as well as the 2-PL model (see online supplemantal material, Table 3); imposes a stricter structure on the items (e.g., equally discriminating items), which is desirable, given our goal of developing a short-form; and has clearer guidelines for evaluating the absolute fit of the model. Despite our primary reliance upon a Rasch or 1PL IRT approach, we also used a 2-PL model to help select items for the short form (for a similar approach, see Nguyen, Han, Kim, & Chan, 2014). Specifically, we used to obtain item discrimination values, which are item slopes, and conceptually may be understood as the extent to which the probability of response relates to the construct being measured. This implies that a more discriminating item is the one for which the probabilities of responses are estimated apart from each other for the respondents at different levels of the construct.

Ordinal modeling. DRI-R items are scored on a 7-point Likert scale. Rather than treat item responses as continuous (as is generally done in confirmatory factor analysis), we used an ordinal item response modeling approach to preserve the ordinal nature of responses. Specifically, we used Masters's (1982) partial credit model (PCM; which uses adjacent-category logit link; Agresti, 2002)—with the marginal maximum likelihood (MML; Bock & Aitkin, 1981) estimation. Although little data were missing in this study, MML estimation handles missing data well (Skrondal & Rabe-Hesketh, 2004). We chose the PCM principally because it fits the data significantly better than the closely related, but simpler model—$\chi^2(40) = 191$, $p < .001$)—the rating scale model (Andrich, 1978). Details on the PCM are provided in the online supplemental material.

Because estimates of the threshold parameters are not necessarily ordered in the PCM model, they do not necessarily reflect the difficulty of achieving a category point. To provide estimates that can be directly interpreted as difficulty parameters, we also present cumulative thresholds in (a) the Wright Map (described below), and (b) the cumulative probability graphs for each of the DRI-SF items (in the online supplemental material).

Multidimensional modeling. The DRI-R consists of three correlated factors. To account for this, we used a three-dimensional version of the PCM rather than a unidimensional version—specifically, we used the multidimensional random coefficient multinomial logit model (also known as the multidimensional Rasch model) proposed by Adams, Wilson, and Wang (1997).

To assess the global fit of this three-dimensional model, we primarily relied on item fit statistics, which are commonly used to evaluate the appropriateness of the Rasch model (Wright & Panchapakesan, 1969; Wright, 1977; Wright & Masters, 1982; Wu, 1997; Wu & Adams, 2013). Item fit statistics describe the discrepancy between observed- and theoretical-item characteristic curves (Wu & Adams, 2013). By convention, "item fit" values between 0.75 and 1.33 are acceptable (Adams & Khoo, 1996).

Tests of cluster (officer) independence. In developing the DRI-R short form, we sought to maximize the Level 1 variance (variance associated with the offender–officer relationship; target variance) and minimize the Level 2 variance (between-officer variance; nuisance variance). To do so, we used a multilevel IRT approach to identify and eliminate items that were heavily influenced by officer traits.

In contrast with other Aim 1 analyses (which use the "combined" dataset), these multilevel analyses focused on the "main" dataset, in which 359 probationers were supervised by 96 probation officers and the average number of probationers per officer was 3.7 (*SD* = 5.4, range = 1–30). To account for probationer response dependence induced by officers, we chose a random-effects (i.e., hierarchical linear modeling) approach because (a) it is infeasible to apply a fixed-effects (i.e., by including dummy variables for each officer) approach in this study, given the small number of probationers within each officer, and (b) a random-effects approach is appropriate, given our wish to make inferences about all probation officers from the relevant population (i.e., generalize the findings to officers beyond those in the dataset).

We estimated the two-level proportional odds model for each item (i.e., response to each item as a univariate outcome) shown below:

$$y^*_{jk} = \xi^{(2)}_{1k} + \epsilon_{jk} \tag{1}$$

in which y^*_{jk} is the latent response for respondent j nested in probation officer k, $\xi^{(2)}_{1k}$ is the officer-specific random effect, and ϵ_{jk} is a respondent-specific random error and has a logistic distribution. This model assumes that officer-specific random effects ($\xi^{(2)}_{1k}$) are independent and normally distributed with a mean of zero. Please see the online supplemental material for the details of the model (see also Skrondal & Rabe-Hesketh, 2004).

Using this model, we identified items with ignorable cluster dependence. In other words, after fitting the above model to each of the items consecutively, we identified items for which accounting for the officer-specific random-effect was not statistically necessary. Then, to test whether the new item set was independent of officers, we accounted for potential clustering in the data by fitting an ordinal multidimensional Rasch model.

Differential item functioning. Our goal was to ensure that the DRI-R short form was sound in terms of its validity for use across respondent subgroups (e.g., adults and juveniles; people with- and without-mental illness). We used differential item functioning (DIF) indices obtained using the Rasch model on the combined dataset to identify and eliminate items that violated the assumption of local independence because of bias with respect to these subgroup factors.

In IRT, an item is considered biased if two respondents from two different subgroups have the same level of the latent variable but have different probabilities of answering in the same response category. DIF is defined in Equation 2 below, where the individual item response is represented by *Y*, the latent variable is represented by θ and a person-level independent variable is represented by *Z* (e.g., age group):

$$P(Y = y|\theta, Z = z) \neq P(Y = y|\theta) \tag{2}$$

This equation implies that the value of Z influences the probability of the response, conditional on the latent variable (θ). The existence of such bias signals that respondents cannot be measured fairly by the item. Following the suggestion in Paek and Wilson (2011) on effects sizes (see also Longford, Holland, & Thayer, 1993), we interpret a statistically significant logit difference value less than 0.426 as "negligible," a value between 0.426 and 0.638 as "intermediate," and a value over 0.638 as "large" DIF. Please see the online supplemental material for the detail on DIF.

External validation analyses for the short form. After following the steps above to select items for the short DRI-R (DRI-SF), we validated the short form in two different ways. First, using the main dataset, we conducted regression analyses to test whether the short version differed from the long version in its: (a) association with probationer ratings of satisfaction in their relationship with their officer; (b) ability to distinguish between specialty and traditional probation supervision programs (using the site variable in the largest dataset); and (c) ability to predict time to rearrest (using Cox survival analysis with robust standard errors because probationers' had varying follow-up periods). For inferential accuracy, we accounted for potential clustering of probationers within officers using random-effects modeling (by doing hierarchical linear regression; see the Tests of Cluster (Officer) Independence section).

Second, using the provider data, we performed an IRT analysis on the DRI-SF items. Unlike the other data sets, the provider dataset was not used to develop the short form. Moreover, unlike the other data sets (where respondents rated relationships with supervising officers), respondents rated relationships with their mandated treatment providers in this dataset. Thus, comparing IRT results of the combined sample with the provider sample may not only help validate the short form in a new sample, but also shed light on its generalizability to a different type of dual role relationship.

MPlus 8 (Muthén, & Muthén, 1998-2017), ConQuest 3 (Adams, Wu, & Wilson, 2012), and Stata 12 (StataCorp, 2011) were used for the estimation of the models presented in the paper.

Results

In this section, we describe the results of IRT analyses that were used to develop a short form of the DRI-R that both focused on relationship variance and was valid across respondent subgroups (Aim 1). We then present the validity of the resulting nine-item short form, using data from both the development- and generalization data sets (Aim 2).

Aim 1: IRT-Based Development of DRI-R Short Form

Baseline analysis of original DRI-R. We began with IRT analyses of the original, 30-item DRI-R to characterize its psychometric properties as a baseline of comparison with those of the short form. DRI-R items and item numbers are provided in Table 1 of the online supplemental material. First, we examined each item's weighted mean square fit statistic, that is, the discrepancy between observed- and theoretical-item characteristic curves. Only three of the DRI-R's items fell outside the acceptable range (Items 3, 6, and 18). Item analysis summary is shown in Table 2 and item parameters and fit statistics are shown in Table 3.

Second, we examined reliability. Marginal reliabilities for all three DRI-R dimensions were ≥.84 and Cronbach's alpha for DRI-R total scores was also high (α = .96; see Table 2).

Third, we tested whether the three-dimensional (i.e., three-factor) model fits significantly better than the unidimensional (single-factor) model, using a Rasch framework. In keeping with results obtained by Skeem et al. (2007) via confirmatory factor analyses, we found that the three-dimensional model fit significantly better than the unidimensional model, χ^2 (5) = 654, $p <$ .001—and that correlations among dimensions were high (absolute values above .80). For the three-factor model, the standardized root-mean-square residual (SRMR) was .06; the root-mean-square error of approximation (RMSEA) was .06; and the comparative fit index (CFI) was 0.96, indicating good overall fit (Hu & Bentler, 1999). For the unidimensional model, the SRMR, RMSEA, and CFI were .07,.03, and .95, respectively.

Development of the short DRI-SF. We used three steps to select the nine items of the DRI-R short form (DRI-SF) from the original scale's 30 items, without compromising the reliability or validity of the instrument. First, we used the combined dataset to flag items for selection when they had relatively high IRT discrimination values (indicating the high correlation between the probability of response in a particular category and the attribute being measured) and made relatively high contributions to scale reliability. Second, we used the main dataset to flag the resulting pool of items for deletion when they exhibited relatively high dependence on clusters, with clusters being probation officers. Third, given the pool of items that survived the first two steps, we identified those that exhibited relatively high differential item functioning (DIF) with respect to mental health status (with-/without mental illness) and age group (juvenile/adult). When items with high DIF were identified, we replaced them with next "best" item (judged by criteria for from step one). We outline the results of each step next. Details for each step are provided in Table 2 of the online supplemental material.

For Step 1, we used the IRT approach described earlier to identify nine items (three per factor) with the highest discrimination values. Note that high discrimination was not the only selection criterion—we also considered the content of the item to avoid redundancy or measurement of a narrow portion of the original construct (Putnam & Rothbart, 2006). All nine of the items with high discrimination values were also found to contribute to the reliability for their respective factors (i.e., would result in relatively high decrements in Cronbach's alpha, if deleted). This process resulted in a set of nine "top" items (original Items 8, 9, 11, 14, 17, 22, 27, 28 and 29). Given that some of these items might perform poorly in Steps 2 and 3, we also identified a set of nine "replacement" items that performed equally well at Step 1 (original Items 8, 9, 11, 14, 16, 21, 22, 27 and 28).

For Step 2, we used the multilevel test of independence to identify items with a relatively high independence of clusters (i.e., probation officers). The goal was to assess relationship quality as freely as possible from officer traits. We first estimated officer-level random-effects for each of the 30 DRI-R items as individual outcomes, and found that Items 14 (toughness), 11 and 27 (trust), and 28 and 17 (caring-fairness) had statistically significant ($p <$ .05) cluster-level variance—indicating that probationers' responses depended on the officer.

Because item-level cluster independence does not always translate to scale-level cluster-independence, we also estimated officer-level random effects at the scale level. For the 9 "top" item set resulting from Step 1 (see above), we found a significant cluster dependence: 21%, 12%, and 13% of variance in the toughness, trust, and caring-fairness dimensions, respectively, were explained by the officer. For the original DRI-R, there was less cluster dependence (8%, 5%, and 6% of variance in toughness, trust, and caring-fairness, respectively, was explained by officers).

To arrive at a DRI-SF with low cluster dependence, we replaced items with significant cluster dependence (in the "top" item set from Step 1) with similarly qualifying candidate items (in the "replacement" item set from Step 1). Specifically, Item 14 was replaced by 24 (toughness), 11 and 27 were replaced by 2 and 26 (trust), and #28 was replaced by #29 (caring-fairness). Reanalysis of the resulting nine-item DRI-SF indicated no significant cluster-level variance (2.0%, 0.4%, and 0.9% of variance in toughness, trust, and caring-fairness, respectively, was explained by officers)—so probationers' responses were not heavily dependent upon probation officers. In other words, the between-officer variance was minimal, and hence the measure was relatively "free" of officer-specific traits.

For Step 3, we used the analyses described earlier to test for differential item functioning (indicating items function in different ways for different subgroups), based on (a) respondents' mental health status (i.e., with/without mental illness), (b) age group (i.e., adolescent vs. adult). The pool of nine items resulting from step two were used for analysis. Results indicated problematic DIF for only one item: Item 24 was moderately biased against adolescent offenders (DIF = 0.59), and was replaced with Item 25. The final items of the DRI-SF are shown in Table 1.

Characteristics of the short DRI-SF. Having selected nine items for the short form, we used the combined dataset to examine the characteristics of the DRI-SF. First, we found that cluster-level variance remained nonsignificant. As shown in Table 3 of the online supplemental material, there was no significant difference between nine-item models that did and did not include clustering across dimensions, less than 1% of the variance was explained by officers.

Second, we checked the spread of estimated item parameters for the DRI-SF to ensure that they represented the spread of those for the original DRI-R. In Figure 1, we show the Wright map (see Wilson & Gochyyev, 2013) from the IRT model for the DRI-R—with items for the DRI-SF shown in rectangles—to visually represent the relationship between person estimates and item difficulties by placing persons and items on a common logit scale. The left side of the figure shows the estimated distribution of respondents across each DRI-R factor, with the highest estimated (i.e., highest scoring) respondents at the top. The right side shows the estimated distribution of items, with the most "difficult" (i.e., least endorsed) items at the top. Given that items have seven ordinal response categories, item 8.6 in Figure 1 represents the 6th threshold for the 8th item. When the respondent and the item have the same logit or

Table 1
Items of the Dual-Role Relationship Inventory—Revised, Short Form (DRI-SF)

Factor	DRI-SF item #	DRI-R item #	Item
Caring-fairness	1	16	____(name) treats me fairly.
Caring-fairness	2	21	____(name) considers my views.
Caring-fairness	3	29	____(name) takes my needs into account.
Trust	4	2	I feel free to discuss the things that worry me with ____(name).
Trust	5	8	I feel safe enough to be honest with ____(name).
Trust	6	26	____(name) knows that he/she can trust me.
Toughness	7	9	____(name) talks down to me.
Toughness	8	22	I feel that ____(name) is looking to punish me.
Toughness	9	25	____(name) expects me to do all the work alone & doesn't provide enough help.

Note. DRI-R = Dual-Role Relationship Inventory—Revised.

location, the respondent has an estimated 50% probability of endorsing the item: When the respondent is above the item, the probability is higher; when the respondent is below, the probability is lower (Wilson & Gochyyev, 2013). Figure 1 chiefly indicates that the subset of DRI-SF items (shown inside rectangles) are spread across the entire range of parameters for the original DRI-R items. Thus, the DRI-SF adequately represents the range of item difficulties for the full 30-item DRI-R. Notably, Figure 1 also indicates that the DRI-SF—like many surveys with Likert-scale items— does a relatively poor job of assessing relationship quality at very high levels of the construct, as there are often respondents who will respond "always" on most or all items. We will elaborate on the implications of this in the Discussion section.

Third, we performed an IRT item analysis of the DRI-SF. As shown in the second column of Table 2, internal consistency (Cronbach's alpha) for the DRI-SF is 0.87 and dimension-level reliabilities all exceed 0.80; and all item fit statistics are within the acceptable range. The fact that all item fit statistics are within the acceptable range indicates that assumptions of the Rasch model were met. The SRMR, RMSEA, and CFI for the three-factor model were 0.03, 0.04, and 0.99, respectively—indicating a good overall fit (Hu & Bentler, 1999). For the unidimensional model, these values were 0.06, 0.03, and 0.96, respectively. The multidimensional PCM model fit as good as its 2PL variant—the multidimensional generalized partial credit model (Muraki, 1992), as shown Table 3 of the online supplemental material.

Table 2
Item Analysis Results for the DRI-R, DRI-SF, and DRI-SF on the Validation ("Provider") Sample

Variable	DRI-R	DRI-SF	DRI-SF ("provider")
Sample size	690	690	125
Number of items	30	9	9
Missing data (%)	3.7%	2.7%	0%
EAP reliability			
Caring-Fairness	.95	.86	.84
Trust	.91	.83	.83
Toughness	.84	.81	.73
Cronbach's alpha	.96	.87	.90
# of items that do not fit the model	3 items	0 items	0 items
# of item categories (steps) that do not fit the model	0 steps	0 steps	0 steps

Note. DRI-R = Dual-Role Relationship Inventory—Revised; DRI-SF = Dual-Role Relationship Inventory—Revised, Short Form.

As was the case with the original scale, the three-dimensional DRI-SF Rasch model—which is shown in Figure 1 of the online supplemental material—fit the data significantly better than the unidimensional model. Correlations among the DRI-SF factors were high (absolute values above .80)—and not statistically significantly (at 0.05 level) different from correlations among factors for the original DRI-R. Cumulative probability curves for each of the nine DRI-SF items are shown in Figure 2 of the online supplemental material. Item parameters and fit statistics are shown in Table 4.

Fourth, we tested the DRI-SF for differential item functioning based on clients' sex and race. We found that none of items from the DRI-SF are biased against any of the sex (male/female) or race (white/nonwhite) groups. This indicates that the instrument can be used for drawing valid inferences on potential differences across these groups.

Aim 2: Validation of the Short DRI-SF

Having developed the DRI-SF and determined that it has favorable psychometric characteristics, we next validated the short form. Specifically, we used the main dataset to determine whether the short form performed similarly to the original form in (a) distinguishing between groups known to differ in relationship quality (i.e., specialty mental health probation vs. traditional probation), (b) predicting probationers' ratings of satisfaction with their relationship with their probation officers, and (c) predicting recidivism. We then used the provider dataset, which had not been used to develop the short form, to examine whether IRT findings from the DRI-SF replicated in a new sample and generalized from probation officers to treatment providers. We begin by describing the validation analyses based on the main dataset.

Validation based on external outcomes. We tested the predictive equivalence of the DRI-R and DRI-SF by comparing the relationship of the original instrument and the shorter form with external outcomes. As a preliminary step, we determined that the association between mean scores on the DRI-R and DRI-SF was very high (r = .97). We also found that the mean score on the DRI-SF predicted the mean score on the full DRI-R, while ac-

Table 3
Item Difficulty Parameters and Item Fit Statistic for the Dual-Role Relationship Inventory—Revised

Factor	Item label	Item difficulty (*SE*)	Infit statistic (T-statistic)	Outfit statistic (T-statistic)
Caring-Fairness	dri_1	−.80 (.03)	.98 (−.3)	1.06 (1.1)
Trust	dri_2	−.45 (.03)	1.23 (3.7)	1.33 (5.6)
Caring-Fairness	dri_3	−1.31 (.04)	1.43 (5.3)	1.53 (8.5)
Caring-Fairness	dri_4	−1.08 (.03)	.91 (−1.3)	.88 (−2.2)
Caring-Fairness	dri_5	−.74 (.03)	1.13 (1.9)	1.34 (5.5)
Caring-Fairness	dri_6	−1.06 (.04)	1.66 (7.0)	2.58 (19.2)
Caring-Fairness	dri_7	−.69 (.03)	1.12 (1.9)	1.47 (7.5)
Trust	dri_8	−.61 (.03)	1.05 (.9)	1.10 (1.9)
Toughness	dri_9r	−1.51 (.04)	1.11 (1.1)	1.05 (1.0)
Caring-Fairness	dri_10	−1.02 (.03)	1.01 (.2)	1.08 (1.4)
Trust	dri_11	−1.07 (.04)	1.03 (.5)	.94 (−1.1)
Caring-Fairness	dri_12	−.91 (.03)	.98 (−.3)	.87 (−2.5)
Caring-Fairness	dri_13	−.98 (.03)	.91 (−1.4)	.80 (−3.9)
Toughness	dri_14r	−1.22 (.04)	1.17 (1.9)	1.25 (4.2)
Caring-Fairness	dri_15	−.86 (.03)	1.09 (1.5)	1.11 (2.0)
Caring-Fairness	dri_16	−1.35 (.04)	.90 (−1.4)	.59 (−8.9)
Caring-Fairness	dri_17	−.84 (.03)	.71 (−5.1)	.57 (−9.2)
Caring-Fairness	dri_18	−.67 (.03)	1.58 (7.6)	2.14 (15.5)
Caring-Fairness	dri_19	−.92 (.04)	1.20 (2.7)	1.10 (1.7)
Caring-Fairness	dri_20	−1.05 (.03)	.86 (−2.0)	1.02 (.4)
Caring-Fairness	dri_21	−.82 (.03)	.86 (−2.4)	.74 (−5.1)
Toughness	dri_22r	−1.13 (.03)	.93 (−.8)	.80 (−4.0)
Caring-Fairness	dri_23	−1.19 (.04)	1.04 (.5)	.96 (−.8)
Toughness	dri_24r	−1.23 (.04)	1.31 (3.6)	1.83 (12.2)
Toughness	dri_25r	−.98 (.03)	1.13 (1.7)	1.15 (2.6)
Trust	dri_26	−.71 (.03)	1.17 (2.6)	1.23 (3.7)
Trust	dri_27	−.53 (.03)	.82 (−3.0)	.73 (−5.5)
Caring-Fairness	dri_28	−.78 (.03)	.82 (−3.1)	.73 (−5.5)
Caring-Fairness	dri_29	−.77 (.03)	.78 (−3.9)	.68 (−6.7)
Caring-Fairness	dri_30	−1.36 (.04)	1.04 (.5)	.97 (−.5)

counting for clustering by officer: The coefficient was 0.89 (*SE* = 0.01), indicating that for every 1-point increase in the mean of the DRI-SF, the mean of DRI-R is estimated to increase 0.89 points.

Turning to external correlates, we first examined whether the DRI-SF distinguished between known groups of relationships in specialty versus traditional probation as well as the original DRI-R. To do so, we regressed the mean score from either the DRI-R or DRI-SF on probation type (specialty vs. traditional), while accounting for clustering by officer. As expected, specialty and traditional groups obtained significantly different mean scores on both the DRI-R (*M* = 5.95 [*SE* = 0.29] and 4.59 [*SE* = 0.15] for specialty and traditional) and DRI-SF (*M* = 5.95 [*SE* = 0.26] and 4.78 [*SE* = 0.15], specialty and traditional). Moreover, the size of the groups' estimated mean differences in scores was similar for the DRI-R (−1.36 [*SE* = .33]; *p* < .001; Cohen's *d*: 1.11 CI [.57, 1.65]) and DRI-SF (−1.17 [*SE* = 0.30]; *p* < .001; Cohen's *d*: 0.85 CI [0.41, 1.29]), indicating that the short form distinguished between groups as well as the original version.

Second, we assessed the measures' association with probationers' ratings of satisfaction in their relationship with their probation officer. We did so by regressing DRI-SF or DRI-R mean scores on the satisfaction score, accounting for clustering by officer. As expected, both the DRI-R and DRI-SF predicted satisfaction ratings. The regression coefficients of the satisfaction scores were estimated at 0.95 (*SE* = 0.04; *p* < .001) and 1.06 (*SE* = 0.04; *p* < .001) for the DRI-R and DRI-SF, respectively—again indicating that the short form performed as well as the original version.

Third, we compared the utility of the DRI-R and DRI-SF in predicting the time to rearrest. We found that both the DRI-R and DRI-SF significantly predicted time to rearrest. Hazard ratios for the DRI-R (0.78 [*SE* = 0.03]; AUC: 0.63) and DRI-SF (0.79 [*SE* = 0.03]; AUC: 0.63) indicate that a one-point increase in mean DRI-R scores and DRI-SF scores correspond to 22% and 21% decrease in the likelihood of rearrest, respectively. Moreover, there were no statistically significant differences in the sizes of the DRI-R and DRI-SF coefficients.

Generalizability to a provider context. In the "provider" dataset—which was not used in any analyses reported above—participants used the DRI-R to rate their relationship with their mandated treatment providers (not their supervising officers). This allows us to evaluate whether our IRT findings for the DRI-SF can be replicated in a new dataset—and generalize to a different dual role relationship context.

Results of the IRT and reliability analyses of the DRI-SF, based on the provider dataset, are presented in the column of Table 2. Although the sample size is smaller than desirable, all IRT item parameters for the provider dataset were similar (within the margins of errors) to those obtained for the combined dataset; and the reliability of the DRI-SF is also acceptable.

Estimated correlations among DRI-SF factors in the provider dataset were not significantly different from those obtained for the combined dataset. Although estimated variances for each dimension were higher for the provider dataset (given the smaller sample size), the pattern of variances was similar to that of the combined

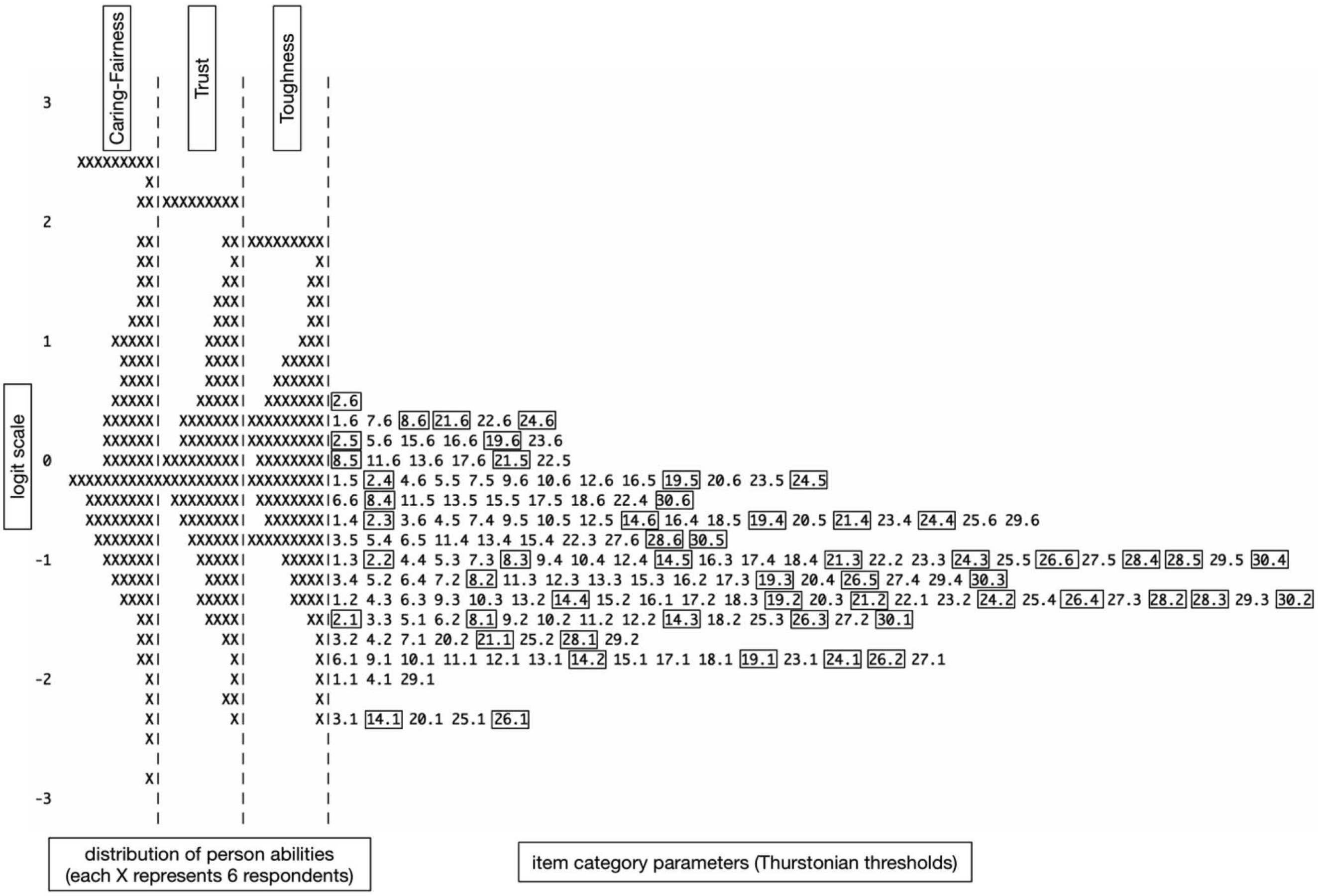

Figure 1. Wright map from the three-dimensional partial credit model using the revised Dual-Role Relationship Inventory (DRI). Items selected for the inclusion in the short-form DRI are shown in rectangles. Items numbered 14, 19, and 24 in the figure measure caring-fairness; items numbered 2, 8, and 21 measure trust; items numbered 26, 28, and 30 measure toughness factors. Horizontal spread of item thresholds is simply a result of multiple thresholds being estimated at equal logit values.

dataset. In fact, there were no significant differences between the combined- and provider-data sets in estimated means of latent variables across all three dimensions. Together, this suggests that the factor structure of the DRI-SF is also consistent across different contexts (i.e., offender-officer and offender-provider).

Discussion

The quality of a client's relationship with a provider appears to influence a broad range of therapeutic outcomes. In traditional treatment contexts, intensive study and understanding of the therapeutic alliance has been fostered by the availability of short, well-validated measures of this construct. In mandated- and "assertive" treatment contexts, research on dual role relationship quality and its influence on outcomes has just begun. Early research suggested that "firm, fair, and caring" relationships were an essential element of effective correctional treatment programs, but relationships were largely assessed via ad hoc, group-based observer ratings (see Dowden & Andrews, 2004). The DRI-R (Skeem et al., 2007) formally operationalizes provider–client relationship quality in mandated treatment, and is well-validated. However, resources are scarce and assessment tools are rarely used in front-line clinical and justice settings where dual role relationships are most common.

The chief purpose of this study was to develop a short form of the DRI-R to promote intensive study of dual role relationship quality and its impact on a range of outcomes. We used multidimensional IRT to reduce the 30-item DRI-R to a nine-item short form, the DRI-SF. Results indicate that the DRI-SF measures all three original dimensions of relationship quality—caring and fairness, trust, and toughness—with negligible changes in reliability. The DRI-SF correlates very strongly with the DRI-R (r = .97), fully represents the DRI-R range of item difficulties (see Figure 1) and reproduces the DRI-R's strength and pattern of correlations with theoretically relevant external covariates (groups known to differ in relationship quality, ratings of relationship satisfaction, and future rearrest). Moreover, the short form does not compromise much on the granularity of scores (which range from 0 to 54; sufficient for differentiating relationships in most settings). A subjective comparison of the content of the DRI-R and DRI-SF suggests that the short form is somewhat more specific in assessing caring-fairness, that is, the short form focuses on fairness whereas

Table 4
Item Difficulty Parameters and Item Fit Statistic for DRI-SF

Factor	DRI-SF item #	DRI-R item #	Item difficulty (*SE*)	Infit statistic (T-statistic)	Outfit statistic (T-statistic)
Caring-Fairness	1	16	−1.92 (.04)	1.08 (1.1)	.89 (−2.1)
Caring-Fairness	2	21	−1.17 (.04)	1.04 (.6)	.85 (−2.8)
Caring-Fairness	3	29	−1.10 (.04)	1.03 (.4)	1.05 (.8)
Trust	4	2	−.39 (.03)	1.05 (1.0)	1.04 (.7)
Trust	5	8	−.53 (.03)	.98 (−.3)	.92 (−1.5)
Trust	6	26	−.62 (.03)	1.08 (1.3)	1.07 (1.2)
Toughness	7	9	−1.52 (.04)	1.11 (1.1)	.82 (−3.5)
Toughness	8	22	−1.13 (.03)	.96 (−.6)	.89 (−2.0)
Toughness	9	25	−.99 (.03)	1.03 (.4)	.88 (−2.2)

Note. DRI-R = Dual-Role Relationship Inventory—Revised; DRI-SF = Dual-Role Relationship Inventory—Revised, Short Form.

the original, much longer scale also includes items that focus directly on the client-provider bond. Subjectively, the remaining two scales of the DRI-SF—trust and toughness—seem comparable to the original scale (which also used relatively few items to assess both constructs). The psychometric properties of the DRI-SF also were replicated in an independent sample that was not used to develop the measure.

Although we demonstrated that the DRI-SF psychometric properties are strong and robust across samples, a reasonable critique is that these samples did not complete the actual nine-item short form. Instead, participants completed the 30-item DRI-R; and we assumed their responses would have been the same if they had completed the nine-item DRI-SF instead. This is a virtually inevitable but important issue in short-form development (Smith, McCarthy, & Anderson, 2000). Thus, the present results need to be replicated in a future study in which the nine-item DRI-SF is directly administered. Counterbalancing this study limitation is our use of a statistical approach that is useful for constructing short forms—an ordinal multidimensional Rasch model (which fit the data as well as the more sophisticated two-parameter logistic model). In the Rasch tradition, we specifically focused on item fit and on ensuring that each item's contribution to the scale was comparable and equivalent (Wright & Masters, 1982).

Perhaps more importantly, we also used statistical techniques to refine the DRI-R in the process of shortening the scale. First, we estimated the dependence of clients' responses to each item on their service providers and then selected items for the short form that manifested minimal dependence. Unlike the DRI-R, DRI-SF scores show no significant "clustering" by provider. The short form has been optimized to cleanly assess the quality of client–provider relationships, while minimizing construct-irrelevant variance associated with provider traits (that can be assessed separately) by selecting items that largely were independent of officer characteristics. Thus, a given service provider can be expected to obtain different DRI-SF scores with different clients—reflecting variance in the quality of her relationships with those clients. Beyond focusing more cleanly on relationship quality, the DRI-SF may also be easier to apply in research (obviating the need for multilevel analyses).

Second, we assessed each item for DIF, or differential item functioning, based on a client's age group (juvenile vs. adult) and mental health status (with/without mental illness). We then selected items for the short form that were unbiased with respect to these characteristics. So, for example, a juvenile- and adult-client with the same latent level of relationship quality have the same probability of endorsing each item in the same response category. Importantly, the DRI-SF is applicable to both juveniles and adults involved in the justice system—whether they experience mental illness or not. We also tested the DRI-SF and found no evidence of differential item functioning by client sex or race. Practitioners and researchers can use the measure with confidence across many client populations to assess dual role relationship quality.

We found a potential "ceiling" effect for the DRI-SF—that is, scores tend to be quite high and skewed. Marked ceiling effects can be problematic. This is the case, for example, if one is interested in measuring the change in dual role relationship quality over time or in using norm-referenced interpretations of total scores. Here, the question is "how big a ceiling effect is too big?" In the context of health surveys, McHorney and Tarlov (1995; see also de Vet, Terwee, Mokkink, & Knol, 2011) defined a small ceiling effects as 1–15% of respondents obtaining the highest score and moderate ceiling effects as >15% of respondents obtaining the highest score. Given this nomenclature, ceiling effects on the DRI-SF are small: 13.8% of respondents obtained the highest score.

Intuitively, one could try to reduce ceiling effects either by adding a category between very often and always (e.g., "almost always"; Lambert et al., 1996) to increase the granularity of the upper end of the scale; or by collapsing categories at the lower end of the scale to reduce the granularity of categories that are not being utilized effectively (Hatcher & Gillaspy, 2006). In our view, this is unlikely to help because a small proportion of "extreme respondents" tend to endorse the top category of Likert-type scales, regardless of the construct being assessed (see Austin, Deary, & Egan, 2006; Eid & Rauber, 2000; Naemi, Beal, & Payne, 2009).

On balance, we believe the small ceiling effects we observed for the DRI-SF are unlikely to be unduly problematic. First, the DRI-SF related in a theoretically coherent manner with a range of criterion variables, suggesting that ceiling effects do not compromise construct validity. Second, these small ceiling effects for the DRI-SF appear to be consistent with those observed for the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989), which is a well-validated measure of the therapeutic alliance that has been shown to capture change over time. Specifically, the average DRI-SF score in our combined dataset is 5.48 on a 7-point scale, which is remarkably similar to average WAI short form scores of 5.02, 5.88 (Hatcher & Gillaspy, 2006) and 5.60 (Falkenström, Hatcher, Skjulsvik, Larsson, & Holmqvist, 2015; see also Hall et al., 2012) on a 7-point scale.

Finally, we confirmed that the psychometric features of the DRI-SF hold across different types of service providers in mandated treatment. Specifically, we demonstrated that the DRI-SF functions well in assessing both officer-client and treatment provider-client relationships. Again, this provides practitioners and researchers with confidence about the DRI-SF's measurement invariance across these provider populations in capturing the relationship quality.

The most obvious advantage of the nine-item DRI-SF over the 30-item DRI-R is a substantially shorter administration time—which is beneficial for both respondents and those that administer and score the instrument. But the DRI-SF also isolates relationship variance, generalizes across several client groups, and replicates across key provider groups. We hope the DRI-SF is used to assess dual role relationship quality across a variety of mandated treatment contexts—from juveniles being treated in the justice system to clients with serious mental illness receiving "assertive" community-based treatment—and that the DRI-SF helps improve understanding of the effect of provider–client relationships on a variety outcomes.

References

Adams, R. J., & Khoo, S.-T. (1996). Quest [Computer Program]. Melbourne, Australia: ACER Press.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23. http://dx.doi.org/10.1177/0146621697211001

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). ACER ConQuest 3.0 [Computer Program]. Melbourne, Australia: ACER Press.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley. http://dx.doi.org/10.1002/0471249688

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573. http://dx.doi.org/10.1007/BF02293814

Angell, B., & Mahoney, C. (2007). Reconceptualizing the case management relationship in intensive treatment: A study of staff perceptions and experiences. *Administration and Policy in Mental Health and Mental Health Services Research, 34,* 172–188. http://dx.doi.org/10.1007/s10488-006-0094-7

Angell, B., Mahoney, C. A., & Martinez, N. I. (2006). Promoting adherence in assertive community treatment. *Social Service Review, 80,* 485–526. http://dx.doi.org/10.1086/505287

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40,* 1235–1245. http://dx.doi.org/10.1016/j.paid.2005.10.018

Benjamin, L. S., Rothweiler, J. C., & Critchfield, K. L. (2006). The use of structural analysis of social behavior (SASB) as an assessment tool. *Annual Review of Clinical Psychology, 2,* 83–109. http://dx.doi.org/10.1146/annurev.clinpsy.2.022305.095337

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. http://dx.doi.org/10.1007/BF02293801

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Burns, T. (2016). Compulsion in community mental health care: Historical developments and current provisions. In A. Molodynski, J. Rugkasa, & T. Burns (Eds.), *Coercion in community mental health care: International perspectives* (pp. 13–22). New York, NY: Oxford University Press.

Burns, T., Yeeles, K., Molodynski, A., Nightingale, H., Vazquez-Montes, M., Sheehan, K., & Linsell, L. (2011). Pressures to adhere to treatment ("leverage") in English mental healthcare. *The British Journal of Psychiatry, 199,* 145–150. http://dx.doi.org/10.1192/bjp.bp.110.086827

Center for Behavioral Health Statistics and Quality (CBHSQ). (2016). *2015 National Survey on Drug Use and Health Public Use File Codebook*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist-Revised (PCL: SV): An item response theory analysis. *Psychological Assessment, 11,* 3–13. http://dx.doi.org/10.1037/1040-3590.11.1.3

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in Medicine: A practical guide*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511996214

Dowden, C., & Andrews, D. A. (2004). The importance of staff practice in delivering effective correctional treatment: A meta-analytic review of core correctional practice. *International Journal of Offender Therapy and Comparative Criminology, 48,* 203–214. http://dx.doi.org/10.1177/0306624X03257765

Duncan, B. L., & Miller, S. D. (1999). *Working Alliance Theory of Change Inventory (WATOCI)*. Retrieved from http://www.talkingcure.com/

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16,* 20–30. http://dx.doi.org/10.1027//1015-5759.16.1.20

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Falkenström, F., Hatcher, R. L., Skjulsvik, T., Larsson, M. H., & Holmqvist, R. (2015). Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. *Psychological Assessment, 27,* 169–183. http://dx.doi.org/10.1037/pas0000038

Gochyyev, P., & Sabers, D. (2009). Item Analysis. In N. J. Salkind (Ed.), *The Encyclopedia of Research Design* (pp. 642–646). Thousand Oaks, CA: Sage.

Hall, A. M., Ferreira, M. L., Clemson, L., Ferreira, P., Latimer, J., & Maher, C. G. (2012). Assessment of the therapeutic alliance in physical rehabilitation: A RASCH analysis. *Disability and Rehabilitation: An International, Multidisciplinary Journal, 34,* 257–266. http://dx.doi.org/10.3109/09638288.2011.606344

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer. http://dx.doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research, 16,* 12–25. http://dx.doi.org/10.1080/10503300500352500

Henry, W. P., Schacht, T. E., & Strupp, H. H. (1990). Patient and therapist introject, interpersonal process, and differential psychotherapy outcome. *Journal of Consulting and Clinical Psychology, 58,* 768–774. http://dx.doi.org/10.1037/0022-006X.58.6.768

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy, 48,* 9–16. http://dx.doi.org/10.1037/a0022186

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working. Alliance Inventory. *Journal of Counseling Psychology, 36,* 223–233. http://dx.doi.org/10.1037/0022-0167.36.2.223

Howgego, I. M., Yellowlees, P., Owen, C., Meldrum, L., & Dark, F. (2003). The therapeutic alliance: The key to effective patient outcome?

A descriptive review of the evidence in community mental health case management. *The Australian and New Zealand Journal of Psychiatry, 37,* 169–183. http://dx.doi.org/10.1046/j.1440-1614.2003.01131.x

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Kaeble, D., & Glaze, L. (2016, December). *Correctional populations in the United States, 2015* (NCJ 250374). Washington, DC: Bureau of Justice Statistics Bulletin.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33,* 499–518. http://dx.doi.org/10.1177/0146621608327800

Kennealy, P. J., Skeem, J. L., Manchak, S. M., & Eno Louden, J. (2012). Firm, fair, and caring officer-offender relationships protect against supervision failure. *Law and Human Behavior, 36,* 496–505. http://dx.doi.org/10.1037/h0093935

Kiesler, D. (1983). The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions. *Psychological Review, 90,* 185–214. http://dx.doi.org/10.1037/0033-295X.90.3.185

Lambert, M. J. (2013). Outcome in psychotherapy: The past and important advances. *Psychotherapy, 50,* 42–51. http://dx.doi.org/10.1037/a0030682

Lambert, M. J., Hansen, N. B., Umpress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., & Reisinger, C. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ 45.2)*. Wilmington, DE: American Professional Credentialing Services.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Manchak, S. M., Skeem, J. L., Kennealy, P. J., & Louden, J. E. (2014). High-fidelity specialty mental health probation improves officer practices, treatment access, and rule compliance. *Law and Human Behavior, 38,* 450–461. http://dx.doi.org/10.1037/lhb0000076

Manchak, S. M., Skeem, J. L., & Rook, K. S. (2014). Care, control, or both? Characterizing major dimensions of the mandated treatment relationship. *Law and Human Behavior, 38,* 47–57. http://dx.doi.org/10.1037/lhb0000039

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68,* 438–450. http://dx.doi.org/10.1037/0022-006X.68.3.438

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. http://dx.doi.org/10.1007/BF02296272

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 4,* 293–307. http://dx.doi.org/10.1007/BF01593882

Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P., Banks, S., Grisso, T., . . . Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services, 56,* 810–815. http://dx.doi.org/10.1176/appi.ps.56.7.810

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176. http://dx.doi.org/10.1177/014662169201600206

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10,* 121–132. http://dx.doi.org/10.3102/10769986010002121

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77,* 261–286. http://dx.doi.org/10.1111/j.1467-6494.2008.00545.x

Neale, M. S., & Rosenheck, R. A. (2000). Therapeutic limit setting in an assertive community treatment program. *Psychiatric Services, 51,* 499–505. http://dx.doi.org/10.1176/appi.ps.51.4.499

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient: Patient-Centered Outcomes Research, 7,* 23–35. http://dx.doi.org/10.1007/s40271-013-0041-0

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71,* 1023–1046. http://dx.doi.org/10.1177/0013164411400734

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment, 87,* 102–112. http://dx.doi.org/10.1207/s15327752jpa8701_09

Skeem, J. L., Emke-Francis, P., & Louden, J. E. (2006). Probation, mental health, and mandated treatment: *A national survey*. *Criminal Justice and Behavior, 33,* 158–184. http://dx.doi.org/10.1177/0093854805284420

Skeem, J. L., Encandela, J., & Louden, J. E. (2003). Perspectives on probation and mandated mental health treatment in specialized and traditional probation departments. *Behavioral Sciences & the Law, 21,* 429–458. http://dx.doi.org/10.1002/bsl.547

Skeem, J. L., Louden, J. E., Polaschek, D., & Camp, J. (2007). Assessing relationship quality in mandated community treatment: Blending care with control. *Psychological Assessment, 19,* 397–410. http://dx.doi.org/10.1037/1040-3590.19.4.397

Skeem, J. L., & Manchak, S. (2008). Back to the future: From Klockars' model of effective supervision to evidence-based practice in probation. *Journal of Offender Rehabilitation, 47,* 220–247. http://dx.doi.org/10.1080/10509670802134069

Skeem, J. L., & Manchak, S. (2010, October). *Final outcomes of the longitudinal study: "What really works!" for probationers with serious mental illness*. Paper presented at the final meeting of the Macarthur Research Network on Mandated Community Treatment, Tucson, Arizona.

Skeem, J. L., & Manchak, S. M., & Montoya, L. (2017). Effects of specialty mental health probation on public safety outcomes: A matched trial. *Journal of the American Medical Association Psychiatry, 74,* 942–948. http://dx.doi.org/10.1001/jamapsychiatry.2017.1384

Skeem, J. L., Winter, E., Kennealy, P. J., Louden, J. E., & Tatar, J. R., II. (2014). Offenders with mental illness have criminogenic needs, too: Toward recidivism reduction. *Law and Human Behavior, 38,* 212–224. http://dx.doi.org/10.1037/lhb0000054

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC. http://dx.doi.org/10.1201/9780203489437

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102

StataCorp. (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

Trotter, C. (1999). *Working with involuntary clients: A guide to practice*. Thousand Oaks, CA: Sage.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231–251. http://dx.doi.org/10.1037/1082-989X.3.2.231

Vidal, S., Oudekerk, B., Reppucci, N. D., & Woolard, J. L. (2013, March) *Examining the link between perceptions of relationship quality with parole officer and self-reported offending among female youth parolees*.

Paper presented at the Meeting of the American Psychology-Law Society Conference, Portland, OR.

Vidal, S., & Woolard, J. (2017). Youth's perceptions of parental support and parental knowledge as moderators of the association between youth-probation officer relationship and probation non-compliance. *Journal of Youth and Adolescence, 46,* 1452–1471.

Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 1–53). Rotterdam, the Netherlands: Sense. http://dx.doi.org/10.1007/978-94-6209-404-8_1

Wright, B. D. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement, 14,* 97–116. http://dx.doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press Chicago.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48. http://dx.doi.org/10.1177/001316446902900102

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models* (Unpublished Masters thesis). Australia: University of Melbourne.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14,* 339–355.

Received March 7, 2018
Revision received September 27, 2018
Accepted October 3, 2018 ■