Efficiently Assessing Firm, Fair & Caring Relationships:

Short Form of the Dual Role Relationship Inventory

Perman Gochyyev, PhD
Graduate School of Education
University of California, Berkeley

Jennifer Skeem, PhD
School of Social Welfare; Goldman School of Public Policy
University of California, Berkeley

Corresponding author: Perman Gochyyev, University of California, Berkeley.

Address: 3659 Tolman Hall, Berkeley, CA, 94720. Email: perman@berkeley.edu
Abstract

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 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 datasets (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.

**Key words:** therapeutic alliance, dual-role relationship, mandated treatment, item response theory, probation, differential item functioning
Efficiently Assessing Firm, Fair & Caring Relationships:

Short Form of the Dual Role Relationship Inventory

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, Fluckiger & 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, Eno 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 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 (9-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 is valid across professional groups (probation officers and treatment providers). We also test the 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, & 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 U.S. (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 U.S. (NSDUH, 2015), 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 UK 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, Schact & Strupp, 1990; Kiesler, 1983). Based on an assessment of 125 provider-client relationships, Manchak, Skeem & 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., Martin, Garske & Davis, 2000; Horvath, Del Re, Fluckinger, & Symonds, 2011).

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 & Andrews (2004) found that correctional agencies staffed by warm, empathic, respectful and non-blaming 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 have 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 Dual Role Relationship Inventory-Revised (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.

**Dual Role Relationship Inventory-Revised**

The DRI-R was created through a multi-stage 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 & Eno 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 multi-method 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, Eno 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 9-items represents a version that is as short as possible, but no shorter than required to achieve desirable psychometric properties. Specifically, a 9-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 9-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 datasets we used and then describe the IRT and external validation analyses.

Datasets

Four datasets were used to construct and validate the short form of the DRI-R. Combined, these datasets 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: (1) probationers’ baseline ratings of satisfaction in their relationship with their probation officers (rated on a 5-point Likert scale), and (2) time to recidivism, defined as whether or not any arrest for a new offense occurred within a minimum two 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 1117 days for probationers with any arrest and no arrest, respectively).

2. The “juvenile” dataset (Vidal & Woolard, 2015) 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, 2013) includes 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 datasets
(i.e., the main, juvenile, and mixed adult datasets) 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 IRB
approval was required because all four datasets were deidentified.

Analyses

Item response modeling to develop a 9-item short form. We used IRT methods to analyze
the items of the DRI-R and its 9-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 1-parameter
logistic model or 1PL. We do so because, relative to alternatives like a 2-parameter logistic IRT
model (or 2PL), the Rasch model fit the data as well as the 2-PL model (see Online Supplement,
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. 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 marginal maximum likelihood (MML; Bock & Aitkin, 1981) estimation. Although little data were missing in this study, MML estimation handles missing data well (Skrondol & Rabe-Hesketh, 2004). We chose the PCM principally because it fits the data significantly better than the closely related, but simpler model—the rating scale model (Andrich, 1978; RSM). Details on the PCM are provided in the Online Supplement.

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 Supplement).

**Multidimensional modeling.** The DRI-R consists of three correlated factors. To account for this, we used a three-dimensional version of the partial credit model (PCM) rather than a unidimensional version—specifically, we used the multidimensional random coefficient multinominal logit model (also known as the multidimensional Rasch model) proposed by Adams, Wilson, & 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 multi-level 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:

$$y^*_j = \zeta_{1k}^{(2)} + \epsilon_{jk}. $$ (1)
in which \( y^*_jk \) is the latent response for respondent \( j \) nested in probation officer \( k \), \( \zeta_{1k}^{(2)} \) is the officer-specific random effect, and \( \epsilon_{jk} \) is a respondent-specific random error and has a logistic distribution. This model assumes that officer-specific random effects \( \zeta_{1k}^{(2)} \) are independent and normally distributed with a mean of zero. Please see the Online Supplement 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 valid 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 3 below, where the individual item response is represented by \( Y \), the latent variable is represented by \( \theta \) and a person-level independent variable is represented by \( Z \) (e.g., age group):

\[
P(Y = y|\theta, Z = z) \neq P(Y = y|\theta)
\]

This equation implies that the value of \( Z \) influences the probability of the response, conditional on the latent variable \( (\theta) \). The existence of such bias signals that respondents cannot be measured
fairly by the item. Following the suggestion in Paek & Wilson (2011) on effects sizes (see also Longford, Holland, and 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 Supplement 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 re-arrest (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 above, “Tests of cluster (officer) independence”).

Second, using the provider data, we performed an IRT analysis on the DRI-SF items. Unlike the other datasets, the provider dataset was not used to develop the short form. Moreover, unlike the other datasets (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 9-item short form, using data from both the development- and generalization datasets (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 Supplement. First, we examined each item’s weighted mean square fit statistic, i.e., the discrepancy between observed- and theoretical- item characteristic curves. Only 3 of the DRI-R’s items fell outside the acceptable range (items 3, 6, and 18). Item parameters and fit statistics are shown in Table 4 of the Online Supplement.

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, $X^2 (5) = 654, p< .001$—and that correlations among dimensions were high (absolute values above .80). The standardized root-mean-square residual (SRMR) was estimated at .06 and root-mean-square error of approximation (RMSEA) was estimated at .06—both indicating a good overall fit (Hu & Bentler, 1999).
**Development of the short DRI-SF.** We used three steps to select the 9 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 Supplement.

For Step One, we used the IRT approach described earlier to identify 9 items (3 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 α, if deleted). This process resulted in a set of nine “top” items (original item numbers 8, 9, 11, 14, 17, 22, 27, 28 and 29). Given that some of these items might perform poorly in steps two and three, we also identified a set of nine “replacement” items that performed equally well at step
For Step Two, we used the multilevel test of independence to identify items with 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 One (see above), we found 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 One) with similarly qualifying candidate items (in the “replacement” item set from Step One). 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). Re-analysis of the resulting 9-item DRI-SF indicated no significant cluster-level variance—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 Three, 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.

[Insert Table 1 about here]

**Characteristics of the short DRI-SF.** Having selected 9-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 non-significant. As shown in Table 3 of the Online Supplement, there was no significant difference between 9-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 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.

Third, we performed an IRT item analysis of the DRI-SF. As shown in the second column of Table 2, internal consistency (Cronbach’s α) 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 standardized root mean square residual (SRMR) was estimated at .03 and RMSEA was estimated at .04—both indicating a good overall fit (Hu & Bentler, 1999). The multidimensional PCM model fit as good as its 2PL variant—the multidimensional generalized partial credit model (Muraki, 1992; GPCM), as shown Table 3 of the Online Supplement.

As was the case with the original scale, the three-dimensional DRI-SF Rasch model—which is shown in Figure 1 of the Online Supplement—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 the Figure 2 of the Online Supplement. Item parameters and fit statistics are shown in Table 5 of the Online Supplement.

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/non-white) 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 versus 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=0.97). We also found that the mean score on the DRI-SF predicted the mean score on the full DRI-R, while accounting for clustering by
The coefficient was 0.89 (0.01), indicating that for every one 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 vs. 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] & 4.59 [SE=0.15] for specialty and traditional) and DRI-SF (M= 5.95 [SE=0.26] & 4.78 [SE=0.15], specialty & traditional). Moreover, the size of the groups’ estimated mean differences in scores was similar for the DRI-R (-1.36, SE=0.33; Cohen’s d: 1.11) and DRI-SF (-1.17, SE=0.30; Cohen’s d: 0.85), 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 (0.04) and 1.06 (0.04) 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 re-arrest. We found that both the DRI-R and DRI-SF significantly predicted time to re-arrest. Hazard ratios for the DRI-R (0.78, 0.03; AUC: 0.63) and DRI-SF (0.79, 0.03; AUC: 0.63) indicate that a one-point increase in mean DRI-R scores and DRI-SF scores correspond to a 22% and 21% decrease in the likelihood of re-arrest, 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 Column 3 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 dataset. In fact, there were no significant differences between the combined- and provider- datasets 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 9-item short form, the DRI-SF. Results indicate that the DRI-SF measures all three original dimensions of relationship quality—Caring & 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 re-arrest). Moreover, the short form does not compromise much on the granularity of scores (which range from 0-54; 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, i.e., the short form focuses on fairness whereas 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 9-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 9-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 9-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 and 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). 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 multi-level 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.

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 relationship quality.

The most obvious advantage of the 9-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


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


*Psychotherapy, 50*(1), 42–51.


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


Figure 1. Wright Map from the three-dimensional partial credit model using DRI-R. Items selected for the inclusion in DRI-SF are shown in rectangles. Items numbered 14, 19, 24 in the figure measure *Caring-Fairness*; items numbered 2, 8, 21 measure *Trust*; items numbered 26, 28, 30 measure *Toughness* factors. Horizontal spread of item thresholds is simply a result of multiple thresholds being estimated at equal logit values.
Table 1. Items of the Dual Role Relationship Inventory-Revised, Short Form (DRI-SF)

<table>
<thead>
<tr>
<th>Factor</th>
<th>DRI-SF item #</th>
<th>DRI-R item #</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Caring-Fairness]</td>
<td>1</td>
<td>16</td>
<td>(name) treats me fairly.</td>
</tr>
<tr>
<td>[Caring-Fairness]</td>
<td>2</td>
<td>21</td>
<td>(name) considers my views.</td>
</tr>
<tr>
<td>[Caring-Fairness]</td>
<td>3</td>
<td>29</td>
<td>(name) takes my needs into account.</td>
</tr>
<tr>
<td>[Trust]</td>
<td>4</td>
<td>2</td>
<td>I feel free to discuss the things that worry me with (name).</td>
</tr>
<tr>
<td>[Trust]</td>
<td>5</td>
<td>8</td>
<td>I feel safe enough to be honest with (name).</td>
</tr>
<tr>
<td>[Trust]</td>
<td>6</td>
<td>26</td>
<td>(name) knows that he/she can trust me.</td>
</tr>
<tr>
<td>[Toughness]</td>
<td>7</td>
<td>9</td>
<td>(name) talks down to me.</td>
</tr>
<tr>
<td>[Toughness]</td>
<td>8</td>
<td>22</td>
<td>I feel that (name) is looking to punish me.</td>
</tr>
<tr>
<td>[Toughness]</td>
<td>9</td>
<td>25</td>
<td>(name) expects me to do all the work alone &amp; doesn't provide enough help.</td>
</tr>
</tbody>
</table>
Table 2. Item analysis results for the DRI-R (column 1), DRI-SF (column 2), and DRI-SF on the validation (“provider”) sample (column 3).

<table>
<thead>
<tr>
<th></th>
<th>DRI-R</th>
<th>DRI-SF</th>
<th>DRI-SF (“provider”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>690</td>
<td>690</td>
<td>125</td>
</tr>
<tr>
<td>Number of items</td>
<td>30</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Missing data (%)</td>
<td>3.7%</td>
<td>2.7%</td>
<td>0%</td>
</tr>
<tr>
<td>EAP reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caring-Fairness</td>
<td>0.95</td>
<td>0.86</td>
<td>0.84</td>
</tr>
<tr>
<td>Trust</td>
<td>0.91</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Toughness</td>
<td>0.84</td>
<td>0.81</td>
<td>0.73</td>
</tr>
<tr>
<td>Cronbach’s Alpha</td>
<td>0.96</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td># of items that do not fit the model</td>
<td>3 items</td>
<td>0 items</td>
<td>0 items</td>
</tr>
<tr>
<td># of item categories (steps) that do not fit the model</td>
<td>0 steps</td>
<td>0 steps</td>
<td>0 steps</td>
</tr>
</tbody>
</table>